PREDICTING PLURALITY: AN EXAMINATION OF THE EFFECTS OF MORPHOLOGICAL PREDICTABILITY ON THE LEARNING AND REALIZATION OF BOUND MORPHEMES

Darcy Elizabeth Rose

Department of Linguistics

University of Canterbury

This dissertation is submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy in Linguistics

September 2017

ABSTRACT

This thesis examines the learning and production of bound morphemes and how this linguistic behavior is influenced by the contextual predictability of the message those morphemes signal. Using the grammatical category of plurality (e.g. cup ~ cups) as a case study, and treating language as a system of message transmission, it demonstrates that the contextual predictability of a grammatical morpheme is correlated with variation in the learning and realization of that morpheme. Building on previous work examining the influence of contextual predictability on linguistic behavior at other levels of linguistic representation, this thesis suggests that a language user's knowledge of morphemes includes some representation of morphological predictability. This informs the larger question of what constitutes a language user's knowledge of language, and how linguistic behavior varies as a function of that knowledge.

The influence of the contextual predictability of bound morphemes and the messages they signal is evaluated via three studies. These studies use the Rescorla-Wagner model and Message-Oriented Phonology, two frameworks which quantify the amount of information carried by a linguistic unit, to examine how the learning and production of bound morphemes is influenced by two biases that shape communication systems. These communicative biases are: a pressure to accurately transmit messages and a pressure to minimize resource costs.

Study 1 explores the effects of morphological predictability on the learning of plural morphemes in an online artificial language learning experiment. Given multiple cues to the morphological category of plurality, this study shows that the second cue is learned less well when the message of plurality is more predictable, given first cue.

Study 2 investigates gradient realizations of plural marking in a spoken corpus of New Zealand English. Using a measure of contextual predictability based on how often the preceding word occurs before a plural noun, this study shows that plural morphemes with higher predictability in context tend to have more reduced realizations.

After demonstrating in Study 2 that contextual predictability plays a role in morphological reduction, Study 3 uses an online rating task to explore how large the relevant context is over which morphological predictability is calculated, and whether this predictability is accessible through subjective ratings. Using extracted contexts of one or five words from the corpus used in Study 2, judgments about the likelihood of a plural occurring in the given context were solicited from native speakers of English.

While moderately correlated with the corpus measure of predictability used in Study 2, neither of the subjective measures of plural predictability is found to be predictive of plural duration. This finding suggests that while the contextual predictability of morphemes does influence production, either language users may not be able to access fine-grained morphological predictability in an overt task, or this task may not have been able to capture such fine-grained intuitions. Further work is required to determine whether an alternative experimental design might elicit subjective ratings which are predictive of plural durations, enabling exploration of the size of the relevant context.

The above studies demonstrate that at the level of the morpheme, in both learning and production, language users are sensitive to the pressures present in any system of communication, and suggest that communicative biases shape human language at the level of the morpheme. These findings invite further research into the interaction of influences of contextual predictability from multiple levels of linguistic structure, as well as exploration of how morphological systems are shaped over time, and how, cross-linguistically, the biases of accurate message transmission and conserving resource cost are balanced to create effective morphological systems.

ACKNOWLEDGEMENTS

First and foremost, I would like to acknowledge the incredible support I have received from my supervisory team: Beth Hume, Jen Hay, and Florian Jaeger. Beth invited me to move to New Zealand the first time we spoke via Skype, before she had even moved herself. While I was skeptical at the time of another big move, she was persistent, and I am so grateful that I finally gave in and jumped ship to come to New Zealand. Beth has been amazingly supportive and encouraging through the entire PhD journey. Through several topic changes, personal ups and downs, and the plethora of distractions I found to keep myself sane, Beth has always known when to offer a hug, when to come watch a basketball game, and when to offer a stern deadline. Jen has been invaluable as an associate supervisor, keeping me on track when I wandered too far afield, making sure I've looked at things from every angle, and providing unwavering support throughout the entire process. Florian has supported me through statistical challenges and making sure I am aware of the breadth of the relevant literature, as well as being an excellent person to talk to when things need to be kept in perspective.

I would also like to thank Janet Pierrehumbert for valuable input on the study presented in Chapter 2, and Kathleen Currie Hall for her input regarding the discussion of MOP, as well as my two thesis examiners, Hunter Hatfield and Fermín Moscoso del Prado Martín.

I am extremely fortunate to have been surrounded by stimulating, welcoming, inquisitive people throughout my entire linguistics career. I would like to thank the students and faculty at Dartmouth for kindling my love of Linguistics, even though I got a late start, and the students and faculty at Indiana University for broadening my knowledge of Linguistics and providing another supportive environment. In particular, I would like to thank Tim Pulju, Dave Peterson, Dan Dinnsen, Ken de Jong, Robert Botne, Valentyna Filiminova, Vitor Leongue, Beatrice Okelo, Sara Sowers-Wills, and Tom Williams.

I would also like to thank the faculty, students, post-docs, and visitors of my latest Linguistics home, the University of Canterbury and the New Zealand Institute of Language, Brain and Behaviour, for helping me to develop skills and grow as a teacher and researcher, for challenging me to constantly improve, and for making it a pleasure to come to the department every day. To the cute office mates, Ksenia Gnevsheva, Wakayo Mattingly, Xuan Wang, and Jacq Jones, thank you for being both incredibly

silly and incredibly insighful, and for always being ready with a spare blanket. Matthias Heyne, Daniel Bürkle, Ryan Podlubny, Ahmad Haidar, Jiao Dan, Andy Gibson, Jacqui Nokes, Mineko Shirakawa, Julia Zimmermann, Keyi Sun, Daiki Hashimoto, and all the rest, I could not ask for a more supportive place to do my PhD, and I am deeply grateful to all of you.

To all of my friends and family, in New Zealand and abroad, thank you for being here for me and providing the love and support I have needed to make it this far. I could not have done it without you.

To all of the basketball coaches, teammates, and players throughout my university career, as well as my Meetup people, thank you for helping me maintain sanity, and giving me a life outside of the University.

To my parents, Jim and Liz, thank you for your unwavering support and for always encouraging me to push my limits and find new challenges. Mom, thank you for instilling a love of languages and teaching me how important it is to travel, and most recently for helping me through the final push, and knowing that I can swim through anything. Dad, thanks for reminding me I don't need to be perfect, for being my biggest and loudest supporter, and for teaching me how to be physical on the basketball court! To my grandparents, Dot, Wes, Alan, and Betty, thank you for placing such a high value on education, and supporting me through my entire course of study. Thank you also for inspiring me to go on adventures and make new discoveries, even if they don't always turn out as planned. To my brothers, Phil and Wes, you're the best brothers a person could ask for. Thank you for reminding me to be goofy, for being as loud if not louder than Dad, and for having my back no matter what. To Yasir, thank you for daring to date someone in the final stage of their PhD, for keeping me fed, and for never doubting. Finally, thank you to Mad for being there through everything, for being my say anything person, and for being my most inspiring BAB.

***

I would also like to acknowledge the University of Canterbury for funding this research through the Doctoral scholarship and several small research grants from the School of Language, Social, and Political Sciences.

The work reported in Chapter 2 was made possible through the support of a subaward under a grant to Northwestern University from the John Templeton Foundation. The

opinions expressed in this thesis are those of the author and do not necessarily reflect the views of the John Templeton Foundation.

The work reported in Chapters 3 and 4 uses the ONZE corpus. The ONZE data was collected by the Mobile Disc recording Unit of the NZ Broadcasting Service, Rosemary Goodyear, Lesley Evans, members of the NZ English class of the Linguistics Department, University of Canterbury, and members of the ONZE team. The work done by members of the Origins of New Zealand English Project (ONZE) in preparing the data, making transcripts, and obtaining background information is also gratefully acknowledged. The Corpus was created and supported with funding from the following sources: University of Canterbury, Foundation for Research, Science and Technology (the New Zealand Public Good Science Fund), the Royal Society of New Zealand, The New Zealand Lotteries Board Fund, and the Canterbury History Foundation. I am particularly grateful to Robert Fromont for his work programming LaBB-CAT – the ONZE Corpus search engine and interactive interface, and for answering my many complicated queries with a smile from halfway around the world.

Finally, thank you to Kayla Friedman and Malcolm Morgan of the Centre for Sustainable Development, University of Cambridge, UK for producing the Microsoft Word thesis template used to produce this document.

CONTENTS

List of Tables

LIST OF FIGURES

List of Appendices

1 INTRODUCTION

1.1 The learning and realization of bound morphemes

This thesis examines the learning and production of bound morphemes, and how this is influenced by the contextual predictability of the message they signal. Bound morphemes, inflectional morphemes in particular, are one way in which abstract grammatical categories can be expressed in language. This thesis focuses on the grammatical category of plurality (e.g. cup ~ cups), and how the learning and realization of cues to plurality vary in an Artificial Language Learning experiment and a corpus of New Zealand English, respectively. Examining bound morphemes through the lens of the predictability of a given morpheme in a given context will inform the understanding of what knowledge language users have of certain linguistic units. This will increase our understanding of the larger question of what constitutes a language user's knowledge of language, and how linguistic behavior varies as a function of that knowledge, in response to biases related to effective communication.

A language user's knowledge has been shown to include statistical properties, including frequency and *n*gram predictability, of linguistic units such as words, segments, and syntactic structures. This has been shown by research demonstrating that linguistic behavior is influenced by these statistical properties, with parts of the signal being enhanced or reduced in correlation with how predictable they are. Jaeger & Buz (2017) provide a thorough review of work demonstrating reduction at various levels of structure. For example, words that are more predictable tend to be reduced (e.g. Bell, Brenier, Gregory, Girand, & Jurafsky, 2009; Bell, Jurafsky, Fosler-Lussier, Girand, Gildea, 1999; Bell, Jurafsky, Fosler-Lussier, Girand, Gregory, & Gildea, 2003;

Gregory, Raymond, Bell, Fosler-Lussier, & Jurafsky, 1999; Jurafsky, Bell, Fosler-Lussier, Girand, Raymond, 1998; Jurafsky, Bell, Girand, 2002; Jurafsky, Bell, Gregory, & Raymond, 2001; Raymond, Dautricourt, & Hume, 2006; Seyfarth, 2014; for review, see Jaeger & Buz, 2017). Likewise, in learning, cues that are less predictable, or contribute more towards disambiguating the message, tend to be learned before those that are more predictable (Ellis, 2006a, 2006b). Overall, the research investigating the influence of predictability (both contextual and context-free, e.g. frequency or relative frequency) on behavior shows a tendency for more predictable units to be reduced in production, and learned later or with more difficulty. More detail on the breadth of work showing the influence of predictability on behavior is provided in Section 1.2. While there is a significant amount of research on the predictability of words and sounds, less work has been carried out at the level of the bound morpheme (but see Ford & Bresnan, 2015; Frank & Jaeger, 2008; Kurumada & Jaeger, 2015; Norcliffe & Jaeger, 2016; Bybee & Scheibman, 1999; for review, see Jaeger & Buz, 2017).

While there are several proposals for the cause of these patterns of predictability-related reduction, this thesis discusses them through the lens of treating language as a system of message transmission. This is a communicative-based approach, which treats probabilistic reduction as a process which facilitates the successful communication of messages, although other potential explanations for reduction include both production-ease accounts and representational accounts (see Jaeger & Buz, 2017). Treating language as a system of information transfer is discussed more thoroughly in Section 1.3, along with some details on alternative explanations. The amount of *information* contained by any given linguistic unit can be defined in terms of probability (Shannon, 1948). Units which are more predictable contain less information. A unit which signals some meaning that is more predictable from context is less important for successful message transmission. This means that it can be reduced in production, or learned later or more slowly, with comparatively little impact on message transmission (see Jaeger & Buz, 2017). However, these observations raise the question of how fine-grained this quantification of information and variation in behavior is. Is information calculated at every level of linguistic structure? And if so, what is the domain over which it is calculated? One level of linguistic structure which has not been investigated to the same degree as other levels is that of the bound morpheme. The overarching goal of this thesis is to provide more evidence informing the question of what constitutes a language user's knowledge of morphemes. This question will be addressed by more specifically

examining whether the statistical properties of morphemes are tracked in the same way that usage patterns seem to be tracked for segments and words. Two frameworks which examine linguistic behavior in terms of *information* are presented in Section 1.3.1.

One reason for which the influences of morphological predictability may not have been addressed as thoroughly as other levels of linguistic structure is that there is a debate about whether morphemes have some degree of representation that is independent of the words to which they are attached. The evidence for and against morphemes having independent representations of some sort, and more specifically whether the phonetic realizations of morphemes and complex words are influenced by morphological structure, is presented in Section 1.4. Presumably, if morphemes do not have independent representations to at least some extent, the prediction would be that their statistical properties are not tracked independently of whole words. When referring to 'independent representations' for morphemes, this thesis is referring to the idea that somewhere in the knowledge of language that language users have, there is a way to track or represent the statistical properties of bound morphemes, such as the frequency of co-occurrence of morphemes with various surrounding contexts, independently of the words to which these morphemes are attached. This thesis does not make any claims about what the nature of this knowledge is, or the exact form that these representations or statistical properties take. By providing evidence that the statistical properties of morphemes are tracked by language users, this thesis provides evidence in favor of morphemes having some degree of independent representation. The nature of these representations is not discussed here. Three studies are carried out which investigate the influence of predictability on the learning and production of bound morphemes.

The remainder of this chapter discusses influences of predictability on linguistic behavior at all levels of linguistic structure (Section 1.2), provides more information about treating language as a system of message transmission (Section 1.3), and then outlines the debate about whether morphemes have a degree of independent representation (Section 1.4). Section 1.5 presents the research questions and hypotheses, followed by a summary in Section 1.6 of the three studies which address these questions.

1.2 Influences of predictability on linguistic behavior

As discussed above, significant amounts of research show that the realization and learning of linguistic units are influenced by their predictability, with more predictable

units tending to be more reduced at many levels of linguistic representation (e.g. *segment*: Bell et al., 2003; Cohen Priva, 2008, 2012, 2015; Gahl, Jurafsky, & Roland, 2004; Gregory et al., 1999; Raymond, Dautricourt, & Hume, 2006; van Son & Pols, 2003; *syllable*: Aylett & Turk, 2006; *word duration*: Bell et al., 1999, 2003, 2009; Gregory et al., 1999; Jurafsky et al., 1998, 2001, 2002; Seyfarth, 2014; *syntax*: Jaeger, 2010, 2011; Kravtchenko, 2014; Tily & Piantadosi, 2009; Wasow, Jaeger, & Orr, 2011; for review, see Jaeger & Buz, 2017), and units which convey already-predictable messages being learned later (e.g. Dietrich, Klein, & Noyau, 1995; for review, see Bardovi-Harlig, 1999).

At the level of the segment, for example, Gregory et al. (1999) find higher rates of deletion of final obstruents in English words which are more frequent, have higher bigram predictability based on the following word, or have higher conditional trigram predictability based on the preceding and following context. Bell et al. (2003) find that in frequent function words in English, vowels are more likely to be reduced in words which have higher bigram predictability based on the preceding or following word. Finally, Raymond et al. (2006) show that English word-internal coronal stops in the onset are more likely to be deleted when the target word is more predictable given the following word.

At the level of the word, Bell et al. (2009) find that English content words tend to have shorter duration when they are more predictable, based on the following word, while Gregory et al. (1999) find that content word duration is shorter when words are more predictable, based on both the preceding and the following word. For function words, Bell et al. (1999) find that function words are shorter when they have higher trigram predictability, based on either the two previous words or one previous and one following word. Bell et al. (2003) find that high frequency, monosyllabic function words are shorter when more predictable in context, based on both the preceding and following word. This effect is independent of the vowel reduction reported above.

At the level of syntactic reduction, there is a tendency to reduce or omit both function words and referring expressions when they are more predictable. For example, optional *that* in both English complement clauses and relative clauses is more likely to be omitted when the constituent following optional *that* is more predictable given the context (Jaeger, 2010, 2011; Wasow et al., 2011). When a referent (e.g. *Thomas*, *he*) is more predictable given the context, English writers are more likely to use a shorter noun

phrase (e.g. *he*) to refer to that referent (Tily & Piantadosi, 2009). Similarly, subject noun phrases in Russian are more likely to be omitted when they are more predictable given the context (Kravtchenko, 2014).

At the level of the morpheme, there are reported effects of both word-level predictability (e.g. Pluymaekers, Ernestus, & Baayen, 2005a) and the predictability of the morpheme itself (e.g. Bybee & Scheibman, 1999; Cohen, 2014, 2015; Frank & Jaeger, 2008; Kurumada & Jaeger, 2015; Norcliffe & Jaeger, 2016) on morpheme reduction or omission, although there are a limited number of studies. The influence of word-level predictability on the realization of morphemes is similar in nature to the influence of word predictability on segments or syllables within words. For example, Pluymaekers et al. (2005a) find that Dutch affixes tend to be shorter when the Mutual Information[1] of the target word and the following word is higher.

The influence of *morphological* predictability is evidenced through effects on optional contraction (Bybee & Scheibman, 1999; Frank & Jaeger, 2008), optional case marking (Kurumada & Jaeger, 2015; Norcliffe & Jaeger, 2016), and gradient realizations of bound morphemes (Cohen, 2014, 2015). Frank and Jaeger (2008) demonstrate that English *not, have,* and *be* are more likely to appear in their contracted forms (e.g. *I've*) when more predictable given either the preceding or following context. Likewise, while not formulated in terms of predictability, Bybee and Scheibman (1999) show that *don't* is likely to be more reduced in contexts which frequently precede or follow *don't*. Kurumada and Jaeger (2015) find that optional direct object marking in Japanese is more likely to be omitted when the intended case of the word is more predictable given the context. Norcliffe and Jaeger (2016) find in Yucatec Maya that speakers are likely to use a reduced form of a relative clause verb when the meaning it signals is more predictable in context. Finally, gradient effects of morphological predictability are found in Cohen (2014) for English and Cohen (2015) for Russian. English third person singular /s/ is shown to be relatively shorter when the third person category is more probable, while Russian first person and third person agreement suffixes (*-i, -o*) tend to be closer in terms of F1 is when contextual predictability is higher.

---

[1] While Mutual Information is not the same as bigram predictability, it is a measure of the likelihood of two words occurring together, so can still be considered probabilistic reduction.

[2] Jaeger and Buz (2017) group the possible explanations of probabilistic reduction into three main

While the above research shows influences of contextual predictability on production, there is also evidence of predictability influencing learning, specifically with regard to morphemes. This evidence includes infant learning (e.g. Kidd, Piantadosi, & Aslin, 2014), second language acquisition (e.g. Dietrich et al., 1995), morphological learning theory (O'Donnell, 2015), and artificial language learning (Beckner, Pierrehumbert, & Hay, 2017; Carr, Smith, Cornish, & Kirby, 2017; Fedzechkina, Jaeger, & Newport, 2012; Fedzechkina, Newport, & Jaeger, 2017; Finley, 2015; Kurumada & Grimm, 2017; Kirby, Cornish, & Smith, 2008).

For example, in adult second-language acquisition, participants have been shown to use lexical means for expressing tense and aspect before learning to use verbal morphology (e.g. Giacalone Ramat, 1995; Meisel, 1987; for review, see Bardovi-Harlig, 1999). In fact, some studies (e.g. Dietrich et al., 1995) show that without instruction, second language learners may not acquire verbal morphology when non-morphological ways of expressing time have already been learned (e.g. temporal adverbials, prepositional phrases, and calendric reference). In this case, the message conveyed by grammatical marking is already predictable without these grammatical markers, so they are not learned.

The influence of predictability has also been shown in infant learning (e.g. Kidd, Piantadosi, & Aslin, 2014), in what is described as a 'goldilocks' effect. When listening to sequences of sounds, infants are more likely to continue paying attention if the sequence presents sounds that are moderately predictable, but less likely to pay attention if the sequence is too predictable or too unpredictable. This research suggests that infants allocate their attention in a way that facilitates some new learning. The authors hypothesize that little attention is paid to predictable sequences because those sequences do not provide an opportunity for learning, while the lack of attention to highly unpredictable sequences may indicate that overly complex stimuli are ignored.

O'Donnell (2015) examines morphological learning from a computational perspective, proposing a theory of morphological learning based on making a system which is able to effectively predict future input. This principle of creating an effective system during learning determines which complex words to store as individual items and which to store as whole words. Storage of individual morphemes facilitates recognition or computation of novel words using those morphemes, while storage of whole words facilitates use of those words in the future. In O'Donnell's formulation, with each new

item encountered, the learner infers which strategy is most likely to allow for prediction of novel forms. With affixes, if a given affix has been encountered with a large number of different stems, then words containing that affix are likely to be stored as a combination of stem + affix. On the contrary, if the affix has been encountered with fewer stems, the complex word is likely to be stored as a whole word. In this way, the learner is using predictability to decide how to store complex words. Productive affixes, which typically have high type frequency, are stored as separate units so they can be easily recombined with new stems. Less productive affixes, which typically have low type frequency, and often high token frequency for those types, are stored with the stems as a whole unit, so that they are available the next time that complex word occurs. While not a straight-forward example of probabilistic reduction, this theory of morphological learning describes an efficient system built on probability.

While the above studies show the influence of predictability on learning in natural languages, research shows that likewise, in artificial language learning, learners are sensitive to the distributional properties of affixes (Finley, 2015), the relationship of form and meaning (Beckner et al., 2017; Carr et al., 2016; Kirby et al., 2008), and the semantic predictability of morphological categories (Fedzechkina et al., 2012, 2017; Kurumada & Grimm, 2017).

For example, in learning morphological segmentation, Finley (2015) shows that learners are more successful segmenting suffixes which have high type frequency than those which have low type frequency, even when accounting for token frequency. After learning stem-suffix combinations in training, participants were more likely to correctly segment novel words consisting of novel stems and high frequency affixes than novel stems and low frequency affixes. Because high frequency suffixes occur with all possible stems in training, they are more likely to occur in novel words than low-frequency affixes, which only occur with a small percentage of possible stems during training. Therefore, participants have more success recognizing more predictable words.

The effect of predictability on learning can also be seen in iterated learning experiments, where some degree of structure emerges as artificial languages are passed from one participant to the next. Kirby et al. (2008) show some emergence of morphological structure referring to the shape, color, and movement of objects. Beckner et al. (2017) replicate and build on these findings by increasing the number of ‘evolutionary chains’ and standardizing the size of training sets. Beckner et al. find an emergence of structure corresponding to differences in shape, color, and number, with the strongest evidence of

standardization coming from items which share the same shape. Emerging structure shows the influence of predictability and reduction on learning because the resulting language requires fewer 'morphemes' than the original random language to convey the same messages. While this iterated learning paradigm does not show probabilistic reduction in the traditional sense, it does suggest that linguistic behavior in learning is influenced by predictability, with languages moving in the direction of requiring less linguistic material to encode a range of meanings. This is true especially when there is a clear communicative goal.

Finally, several studies show an effect of predictability on the use of optional morphology in artificial languages. Fedzechkina et al. (2012) find that adult learners of artificial languages which are inefficiently organized for information transfer tend to restructure those languages to make them more efficient. Given verb-final artificial languages with variable constituent order and optional case marking, with equal amounts of case marking on animate and inanimate constituents, the learners produce output with more object marking for animate objects, and more subject marking for inanimate subjects. Because inanimate subjects and animate objects are more atypical, this suggests a preference for structure which is more efficient for information transfer. In terms of predictability, participants are restructuring the languages so that more predictable objects or subjects have a more reduced realization (no case marking), while less predictable objects or subjects have a less reduced realization (case marking). Fedzechkina et al. (2017) build on this study by again using artificial languages with optional object marking. However, one condition has variable word order and one condition has fixed word order. Participants in the variable word order condition use case marking in production more often than participants in the fixed word order condition, demonstrating a tendency to use case marking when the grammatical function of the constituents is not predictable based on word order. Again, this shows reduction when the message (here, grammatical function) is more predictable.

In terms of plural morphology, Kurumada and Grimm (2017) show that participants learning an artificial language are more likely to use optional plural marking when the target is semantically less likely to be plural (animals vs. insects). This shows that when the message of plurality is more predictable, less linguistic material is used, which is a form of reduction.

The evidence presented in this section regarding the influence of predictability on both production and learning suggests that language users have access to details about the predictability of linguistic units in context, not just at the level of the segment and word, but also the level of the morpheme. In the production literature, there is a significant amount of work examining the effects of predictability on the realization of words and segments. However, there is less work examining the realization of morphemes. Most of the work that addresses morphological predictability examines either categorical presence vs. absence of a morpheme (Bybee & Scheibman, 1999; Frank & Jaeger, 2008;) or whether the morpheme occurs in a contraction or in its full form (Kurumada & Jaeger, 2015; Norcliffe & Jaeger, 2016). The few studies that look at gradient realizations of morphemes as a function of morphological predictability use experimental elicitation (Cohen, 2014, 2015; Seyfarth, 2016). This thesis builds on the production literature by examining the influence of morphological predictability on the gradient realization of a bound morpheme, using speech from a corpus. In the learning literature, there is evidence that when the message conveyed by a grammatical marker is more predictable from the surrounding context, that marker is likely to be omitted (Fedzechkina et al., 2012, 2017; Kurumada & Grimm, 2017). However, this thesis adds to that literature by examining the extent to which the learning of a cue to a morphological category (e.g. plurality) is influenced by whether the information signaled by that cue is predictable, given other cues within the same word. This thesis provides further evidence that predictability has an influence on the production and learning of bound morphemes.

As discussed above, there are several proposed reasons for the patterns of reduction of more predictable units. The approach taken in this thesis is a *communicative* approach[2] (see Jaeger & Buz, 2017), in which it is assumed that the goal of the speaker is to communicate effectively. In general, communicative approaches assume two competing biases: one which aims to expend as few resources (e.g. time, energy, cognitive effort) as possible, and another which aims to accurately communicate the desired message. Similar arguments have been made for learning, with units that convey more

[2] Jaeger and Buz (2017) group the possible explanations of probabilistic reduction into three main categories: production ease, communicative, and representational. The studies presented here are potentially compatible with other accounts, but the focus here is on communication.

information being learned earlier, presumably in order to facilitate message transmission. The idea of quantifying language in terms of information, and treating language as a system of message transmission, is explored in the next section.

1.3 Language as a system of message transmission

Under the assumption that one of the primary functions of language is to facilitate the transmission of messages from one language user to another (or to several others), then language is a communication system, and thus subject to biases which shape any system of communication. The ideal communication system should (1) use the shortest possible code (e.g. acoustic signal) which allows for unambiguous encoding of all possible messages, and (2) should allow for the transmission of these messages with a relatively low error rate (Jaeger, 2013; Pate & Goldwater, 2015; Shannon, 1948). There is an additional consideration of conserving resources, which is related to using the shortest possible code, but also takes into account the resources necessary for planning utterances. This is discussed in the ideal talker model (Ferrer-i-Cancho, 2005), but the present discussion focuses on the trade-off between efficient code and low rates of transmission error.

When treating language as a system of message transmission, there is an intended message that needs to be transmitted from the producer (talker) to the perceiver (listener). Because ideas cannot be transmitted directly from talker to listener, the message must be converted into a code (e.g. an acoustic signal). This signal is then transmitted through some channel, after which the listener infers the intended message (see Pate & Goldwater, 2015). It is during this process that the two biases above are important. In order to satisfy the first bias, having the shortest possible code to unambiguously transmit all possible messages, a language should have shorter segments of code (e.g. words or morphemes) for more frequent or more predictable messages. If a message which occurs frequently is encoded by a short word, then the overall length of the code will be shorter. Effects of this sort can be seen in work showing that the length and duration of words tend to be shorter when they are more frequent or on average more predictable (e.g. Zipf, 1949; Piantadosi, Tily, & Gibson, 2011; Seyfarth, 2014). In a system with no noise, this optimal code, where the most frequent words have the shortest signal, will also result in error-free message transmission, because every message has a unique code. This would satisfy the second bias presented above. However, communication rarely, if ever, takes place in an environment with no noise.

As soon as noise is introduced, pieces of the signal can be lost during transmission. Given the shortest possible code, if any piece of the code is lost during transmission, some part of the message is also lost. In order to avoid this, *redundancy* can be introduced into the signal. Redundancy is some amount of code or signal which reinforces the message being sent, but is not strictly necessary to transmit a message in a noise-free environment (Hall et al., submitted, based on Pierce, 1980:292). Introducing redundancy allows messages to be recovered even if pieces of the signal are lost during transmission.

While redundancy helps to improve the probability of successful message transmission, in response to the second bias, reinforcing each part of the message in this way would result in a very long code, going against the first bias. In order to balance these two pressures, the ideal language should introduce redundancy to reinforce only certain parts of the message, chosen either because they are likely to be corrupted by noise, or because they are less predictable from context. But inferring which parts of the message should be reinforced requires more knowledge. Selecting parts of the signal that are likely to be corrupted by noise requires making inferences about the noise in the system. One example of speakers adjusting the signal based on predicted noise is the Lombard effect (Lombard, 1910a, 1910b, 1911), where productions become more intelligible in noisy environments. However, predicting noise is not the only way to select which parts of the signal should be reinforced, and which can be reduced. Another way to select pieces of the signal to reduce is by selecting those which carry more *information*. Shannon (1948) measures the contribution of each piece of the signal (e.g. /b/) using information, which is equal to the negative log of the probability of the segment (e.g. $-\log_2 p(/b/)$) given the context.

Words or segments which carry more information contribute more to disambiguating the intended message (e.g. the word-initial /b/ in *bubble*). This means that in order for the message to be transmitted successfully, it is important that these particular pieces of the signal be successfully transmitted. On the contrary, segments which are more predictable and therefore carry less information are less important to successful message transmission. In order to satisfy the biases of using the shortest possible signal and maintaining a relatively low rate of transmission error, segments or words which have higher information are good targets for adding redundancy, while those with low information are good targets for reduction.

Treating language as a system of information transfer, sensitive to the biases of using the shortest possible signal and maintaining a relatively low rate of transmission error, is one way of explaining the patterns of reduction discussed in Section 1.2. The example below takes the English plural suffix as an example.

In a plural noun, e.g. *dogs*, there are at least two messages being conveyed: the animal 'DOG' and the message of plurality 'PL'. Depending on whether morphemes are stored independently or as part of whole words, these messages may be indexed separately, or as part of one combined message, for example 'more than one dog'. For the purposes of this discussion, plurality is treated as a message that can be separated at least to some extent. However, the same main principles apply if the word is treated as a whole. This example will focus on the message of plurality. The first bias, using the shortest code possible which is unique to the message, would push the /s/ to be very short. The main message competing with PL is that of the singular, so the /s/ has to be at least minimally longer than the singular suffix, /∅/.[3] The second bias, keeping the probability of successful message transmission high, would push the /s/ to be longer, to differentiate it to a greater degree from the singular. Assuming for now that noise is constant, the way these two biases interact is based on the predictability of the message PL. To illustrate this, below are two example sentences where the predictability of the PL message is different.

1. All of the dogs were barking.
2. My neighbor's dogs ran outside.

The plural morpheme /s/ indexes the message of plurality. In (1), the surrounding context makes the message of plurality highly predictable. The expression *all of the* and the verb *were* both indicate the plurality of the noun. This means that the plural morpheme /s/ has high predictability,[4] and therefore low information content. It is not contributing much to identifying the message of plurality. On the contrary, in (2), there are no indications in the surrounding context about whether the noun is singular or plural; the sentence containing the singular, *My neighbor's dog ran outside*, is also licit.

[3] Or, in the whole-word approach, the message most closely competing with *dogs* is *dog*.

[4] This measure of predictability is not the same used traditionally to calculate information. However, the same principles apply. This will be discussed further in Chapter 3.

In this case, the plural morpheme /s/ has low predictability and therefore high information content; it is crucial to understanding the message.

According to the biases outlined above, the /s/ in (1) is a good target for reduction. Most likely, even if /s/ is not produced at all, the message of plurality will be conveyed successfully. In (2), the /s/ is not a good target for reduction, since reducing /s/ significantly lowers the likelihood of successful message transmission. Across all instances of plural /s/, the prediction from treating language as a system of message transmission is that in sentences like (1), /s/ will be reduced due to the influence of the bias of using the shortest code possible, while in sentences like (2) it will not, to satisfy the bias of maintaining low transmission error rates. Two proposals for measuring the contextual predictability of plural /s/ can be found in Chapters 3 and 4.

To summarize, if language is treated as a system of information transfer, there are several key properties that allow for the quantitative analysis of linguistic behavior. The amount of *information* carried by a word or morpheme is determined by how *predictable* that word or morpheme is. The two biases of using the shortest possible signal and maintaining a relatively low rate of transmission error work together to create an effective linguistic system, in which more signal is used for high-information segments, and less signal for low-information segments.

The studies presented in this thesis demonstrate that these principles related to information transfer and effective communication systems are operating at the level of the morpheme, by showing that producers and comprehenders display different behavior depending on the amount of information carried by morphemes. This indicates that language users have access to statistical information about the usage patterns of morphemes. The following section discusses two frameworks which quantify information in language, in both learning and production.

1.3.1 Two frameworks which quantify information in language

A wide range of research in linguistics has used various frameworks which quantify the information carried by linguistic units to predict behavior. This thesis focuses on two frameworks: the associative learning framework, specifically the Rescorla-Wagner model (RW, Rescorla & Wagner, 1972), and Message-Oriented Phonology (Hall et al., submitted). A brief outline of each framework is provided here.

The Rescorla-Wagner model is a formal way of quantifying patterns of associative learning, a process by which participants come to associate certain cues with certain outcomes, and was originally formulated to explain patterns of animal learning (Rescorla, 1969; Rescorla & Wagner, 1972; Wagner, 1969, 1970). Since then, this model, along with some modifications, has been successfully applied to language learning in both children and adults, in first and second language acquisition (e.g. Ellis, 2006a, 2006b; Bardovi-Harlig, 1992; Ramscar, Dye, & McCauley, 2013). In applying this model to learning plural marking, linguistic units are treated as cues and the message of plurality is treated as the outcome. The RW model makes predictions about how well various cues will be learned, depending on their statistical properties within a system. Crucially, the RW model is based on the learner modifying the associative strength of cues based on prediction errors related to prediction of the outcome (successful message transmission), and is not concerned with whether the learner successfully perceives each individual cue. In this thesis, predictions from associative learning will be used to predict behavior in an online artificial language learning experiment where the redundancy of morphological cues varies across conditions (Chapter 2).

Previous work using the RW model to predict linguistic behavior finds that patterns of learning of linguistic units in both first and second language acquisition can be explained by the RW model (Ellis, 2006a, 2006b; Bardovi-Harlig, 1992; Ramscar et al., 2013). The study in Chapter 2 uses the RW model to predict behavior in an online Artificial Language Learning task, where participants are learning two cues to plurality. Across conditions, the different cues have varying levels of redundancy, and thus are expected to be learned with varying degrees of success. Differences across conditions in the rate of selection of the two cues to plurality in the test phase indicate that the learning of multiple cues to plurality is influenced by their redundancy, as predicted by the RW model.

The second framework which will be used is Message-Oriented Phonology (Hall et al., submitted). This theory creates a unified picture of the biases which shape phonological systems, drawing from previous work in both phonology and other domains of linguistics. A crucial assumption of MOP is that the main function of language is not to successfully transmit the identity of each individual phoneme, but rather the overall message. This pressure towards successful message transmission is balanced by the

pressure of keeping resource cost low, a slight reformulation of the two biases presented above. Ideas from MOP will be used in this thesis to examine the influence of the contextual predictability of plurality on the duration of plural /s/ in New Zealand English (Chapters 3,4). Chapter 3 provides more details about the key components of MOP.

While MOP is a new theory, it draws on a large body of research, both within phonology and in other domains of linguistics, which has demonstrated the influence of predictability on the realization of linguistic units (see Section 1.2), using insights from Information Theory (Shannon, 1948) and Bayesian inference (Bayes, 1763; Laplace, 1812).

While MOP focuses primarily on words as messages, it also acknowledges that messages can be above the word level (e.g. social messages) or below the word level (e.g. morphemes). Guy (1996) treats morphological categories (e.g. number, tense), as messages that are conveyed, and suggests that there is no clear evidence that the amount of information carried by a morpheme influences the phonetic realization of that morpheme. However he does hypothesize that there is a pressure on communication systems to maintain clarity in conveying morphological information, and that this could act at the level of acquisition. In addition, recent findings regarding the influence of morphological predictability on the realization of morphemes (e.g. Cohen, 2014, 2015) provide further evidence that the realization of morphemes is shaped by the biases of successful communication and keeping resource cost low. The studies in Chapters 3 and 4 provide further evidence that morphemes are governed by these same properties.

While the information carried by morphemes can be quantified, as shown here, there is an active debate about whether morphemes have independent representations of some nature. The evidence for and against this idea is presented in the following section.

1.4 Regarding the independent representation of morphemes

The assumption that morphemes are ‘psychologically real’ and have some amount of representation that is independent of their base is a controversial topic. As stated in Section 1.1, this thesis does not make claims about the nature of these hypothetical representations, it simply argues that the statistical properties and co-occurrence patterns of morphemes are stored independently of whole-word properties. One area with rich evidence about the representations of morphemes is lexical decision tasks, either with or without priming. Evidence from research in this area suggests that the

degree to which morphemes have independent representations is gradient, and varies based on factors such as the type of affix (inflectional vs. derivational; e.g. Laudanna, Burani, & Cermele, 1994), consistency (e.g. Chateau, Knudsen, & Jared, 2002), and transparency (e.g. Gonnerman, Seidenberg, & Andersen, 2007; Marslen-Wilson, Tyler, Waksler, & Older, 1994). This section first provides some background related to lexical decision tasks and affix priming, then examines acoustic evidence for and against the proposal that morphological structure influences the phonetic realization of morphemes. If morphological structure influences the phonetic realization of morphemes, it would provide further evidence that at least some morphemes have some degree of independent representation.

Lexical decision tasks using both real and non-words, with or without masked primes, provide insight into the status of morphological representations. With no primes, Caramazza, Laudanna, and Romani (1988) show that non-words which bear closer resemblance to real words (by either including a real stem, a real inflectional affix, or both) take longer for participants to reject. This evidence that affixes that occur in real words cause slower response times suggests that there is some representation of these affixes which is activated, even when they occur in a non-word.

In lexical decision tasks with primes, primes with either the same stem or the same affix as the target result in shorter response latencies. This suggests that both the stem and affix have some representation distinct from the representation of the complex word as a whole. Studies using primes which share the same stem as the target have found evidence that lexical decision for complex words is facilitated by a prime with the same base, when the complex word and the base word have a semantically transparent relationship, and that this effect can be gradient. Words which are more semantically transparent show more facilitation (Grainger, Colé, & Segui, 1991; Gonnerman, Seidenberg, & Andersen, 2007; Marslen-Wilson, Tyler, Waksler, & Older, 1994).

However, more relevant to the question of whether bound morphemes have independent representations of some nature, there is evidence that priming with words that share an affix with the target facilitates recognition of words (e.g. Chateau, Knudsen, & Jared, 2002; Dominguez, Alija, Rodgriguez-Ferreiro, & Cuetos, 2010). The evidence here is not entirely consistent, but points to affixes having independent representations. Chateau et al. (2002) show that facilitation occurs when prefixed words are primed by words which share the same prefix, although the effect differs depending on whether the

prefix is ‘high-consistency’ (the sequence of letters that form the prefix almost always serves as that prefix, e.g. UN) or ‘low-consistency’ (the sequence of letters occurs more often at the beginning of monomorphemic words, e.g. DE). For high-consistency prefixes, facilitation occurs regardless of whether the prime has a real or pseudo-prefix, while pseudo-prefixed primes with low-consistency prefixes do not provide facilitation. This suggests that the meaning of the morpheme is activated in some way for words with high-consistency prefixes, or that these prefixes have some kind of independent representation. A similar result is found in Spanish (Dominguez et al., 2010), with both prefixed and pseudo-prefixed words facilitating recognition of targets with the same prefix, when the stimulus onset asynchrony (SOA) is short. However, when the SOA is longer, only true prefixed primes facilitate recognition of prefixed words, while pseudo-prefixed words inhibit recognition. Again, this suggests that the meaning of the prefix is activated by the prime.

Overall, research using affix priming and lexical decision tasks suggests that bound morphemes do, at least to some extent, have independent representations. However, this is not consistent across all bound morphemes. The degree to which morphemes seem to exist independently is influenced by whether they are inflectional or derivational, how consistent they are, and how transparent they are. According to this research, the affixes which are most likely to have independent representations are high-transparency, high-consistency, and inflectional. Plural /s/ in English fits all of these criteria. The following section explores evidence in the acoustic domain about the independence of morphemes.

1.4.1 Acoustic evidence

While the evidence above suggests that at least some morphemes have independent representations of some nature, based on lexical decision tasks with or without primes, this thesis focuses on evidence that the phonetic realization of morphemes and segments is influenced by morphological structure. While an absence of phonetic evidence would not prove that morphemes do not have independent representations, the presence of evidence showing that morphological structure influences phonetic realization strongly suggests that morphemes have independent representations, at least to some extent. In support of the idea that morphological structure is relevant to phonetic detail, and that morphemes have independent representations, there is both acoustic and articulatory evidence, based on differences in the characteristics of both stems and affixes.

However, some of these studies have been criticized for small sample sizes or unbalanced datasets. Some of these studies compare segments which form parts of affixes to those that do not (e.g. the /s/ in *laps* vs. *lapse*), while others look within a given affix to find differences based on decomposability or predictability. The evidence from articulatory studies suggests that morphological structure plays a role, but more work is needed (Cho, 2001; Song, Demuth, Shattuck-Hufnagel, & Ménard, 2013; Mousikou, Strycharczuk, Turk, Rastle, & Scobbie, 2015). In the domain of acoustics, there is evidence both for and against morphological structure influencing phonetic realizations. Findings suggest that morphological structure plays a role in how segments in complex words are produced (e.g. Hay, 2004; Plag, Homann, & Kunter, 2017; Schuppler, van Dommelen, Koreman, & Ernestus, 2012; Seyfarth, 2016; Zimmermann, 2016). However, other researchers have claimed that such effects are due not to the morphological structure of the word, but rather to how much a given segment contributes to identifying the word as a whole (*word information load*; e.g. Hanique & Ernestus, 2012). The former suggest that morphemes are psychologically real entities, while the latter suggest the opposite. The remainder of this section explores the acoustic evidence in more detail, starting with differences between the realization of complex words and monomorphemic words, then examining differences within the realization of complex words that share a given affix.

The first type of evidence regarding the influence of morphological structure on phonetic realizations compares the realization of segments in complex words to the realization of those same segments in monomorphemic words. These studies look at effects on suffixes (Losiewicz, 1992; Walsh & Parker, 1983; Schuppler et al., 2012), prefixes (Baker, Smith, & Hawkins, 2007; Hawkins & Smith, 2001), and stem-final syllables (Sugahara & Turk, 2009).

In the domain of suffixes, Losiewicz (1992) and Walsh and Parker (1983) provide evidence that English word-final coronal stops and fricatives, respectively, are longer when they are grammatical morphemes (past tense or plural) than when they are not. These studies use apparent homophones such as *rapt* vs. *rapped* and *lapse* vs. *laps*, and find that the word-final coronals are longer in the complex words. However, both of these studies have been criticized for having relatively small sample sizes as well as a potential confound of word frequency, with the monomorphemic words being almost always more frequent than the complex words (Hanique & Ernestus, 2012; Hay, 2004;

Plag et al., 2017). Additionally, Schuppler et al. (2012) examine the presence or absence of Dutch word-final /t/ across all content words, and initially find no effect of the morphological status of the /t/. However, they then restrict their dataset to pairs of words which are phonemically identical, but differ in whether the final /t/ is a simple final /t/, or a degeminated stem-final /t/ plus an independent morpheme (e.g. *vind* '[I] find' vs. *vindt* '[he] finds', both pronounced [vɪnt]). In this limited dataset, [t] is more likely to be deleted when it does not have an independent morphological function, a result which is in accordance with that found by Losiewicz. However, Hanique and Ernestus (2012) caution that this finding is also based on a limited dataset.

With regard to prefixes, *mis-* and *dis-* are found to have different relative durations of segments depending on whether they are true prefixes or pseudo-prefixes (e.g. *mistimes* vs. *mistakes*, Baker et al., 2007; Hawkins & Smith, 2001). These studies propose that the differences may be due to the differences in productivity between the two types of sequences, and that they help to reduce the number of possible words by the end of the *mis-/dis-* sequence. It is not clear whether this result is due to effects of morphology or word identification.

Finally, with regard to stem-final syllables, Sugahara and Turk (2009) find that stem-final syllables in complex words with Level II affixes (e.g. –s, –ed, –ing) are longer than corresponding sequences in monomorphemic words. They suggest that this is due to differences in the prosodic structure of complex words, also indicating that morphological structure matters.

In addition to differences between complex and monomorphemic words, studies which compare the realization of either the stem or affix within the set of words containing a given affix provide evidence regarding the influence of morphological structure on phonetic realizations. These studies show an effect of either the relative frequency of a complex word to other words in the paradigm, or the morphological predictability of a morpheme, demonstrating that morphological structure is relevant.

With regard to stem segments in complex words with the same affix, Hay (2004) finds that in English *–ly* words, complex words which are more frequent than other words with the same stem (e.g. *swiftly*) tend to display more /t/ reduction than complex words which are less frequent than other words with the same stem (e.g. *softly*). Hay suggests that this is due to different types of processing being active for words with higher and lower relative frequencies. While both whole-word and decomposition processes are

active for most words, words with higher relative frequencies are more likely to be processed as whole words, while those with lower relative frequencies are more likely to be processed as stem + affix (see also Hay & Baayen, 2005).

With regard to affix segments, Schuppler et al. (2012), find that word-final /t/ in Dutch complex words is more likely to be present in words with higher relative frequency. While these two results go in opposite directions, one showing reduction in words with higher relative frequency and the other in words with lower relative frequency, they both suggest that morphological structure is relevant to phonetic realizations.

While the evidence presented thus far seems to indicate that morphological structure is important to the acoustic realization of morphemes, suggesting that morphemes have independent representations, some researchers argue that the evidence is not conclusive. The principle arguments against morphemes being psychologically independent units, as evidenced by acoustic data, is presented in Hanique and Ernestus (2012), who begin with the idea that if morphological structure is important in processing, then reduction effects that have been shown for whole words should also be seen for morphemes. They address three scenarios in which words tend to be pronounced with reduced phonetic realizations, then review the literature which investigates morphemes under these same three scenarios. After reviewing the literature and reanalyzing several datasets, Hanique and Ernestus conclude that morphological structure is not important in phonetic reduction, and thus that all complex words are stored and processed as complete units.

The three areas where Hanique and Ernestus expect reduction of morphemes, if morphological structure is relevant, are when morphemes are repeated, when a segment plays a more crucial role in identifying a morpheme, and when a word is more morphologically decomposable. For the first point, they cite one study which found no effect of repetition (Viebahn, Ernestus, & McQueen, 2012), while for the third, they claim that evidence is too inconsistent to be conclusive (Bürki, Ernestus, Gendrot, Fougeron, & Frauenfelder, 2011; Hanique, Ernestus, & Schuppler, 2013; Hay, 2004; Schuppler et al., 2012). It is the second area, the contribution a segment makes to identifying the morpheme, that is particularly relevant to this thesis, and is discussed further. Chapters 3 and 4 expand on the claims presented here and how they are addressed in the present studies.

The contribution a segment makes to identifying the morpheme is relevant to both the difference between complex words and monomorphemic words, and variation within a

given morpheme. Hanique and Ernestus propose that if morphological structure is important, single segment affixes should display less reduction than segments at the end of longer affixes or segments in monomorphemic words. This type of effect is shown in the above studies (Losiewicz, 1992; Schuppler et al., 2012; Walsh & Parker, 1983), but Hanique and Ernestus point out that for each of these studies, there is either a relatively small sample size, the results could be due to factors other than morphological structure, or there are other problems with the dataset.

More recently, two large corpus studies of New Zealand and American English find that there are systematic differences between the duration of different types of word-final /s/ in English, including differences between morphological /s/ and non-morphological /s/ (Plag et al., 2017; Zimmermann, 2016). Unlike the findings of Walsh and Parker (1983), these two studies find that non-morphemic /s/ has the longest duration, followed by morphemic /s/ such as plural and 3rd person singular, and finally, clitic /s/ tends to have the shortest duration. Although these findings do not agree with a probabilistic reduction hypothesis, they do provide stronger evidence that morphological structure is relevant to phonetic realizations, using a large dataset and controlling for confounding factors.

With regard to variation within affixes, Hanique and Ernestus again claim there is no definitive evidence in favor of morphemes having independent representations of some nature, despite research which seems to show such an influence (e.g. Pluymaekers, Ernestus, Baayen, & Booij, 2010). They propose that any effects on production which have previously been attributed to morphological structure can be attributed to other factors such as word frequency or *word information load* (van Son and Pols, 2003), and that morphemes emerge as categories based on complex words which share phonetic and semantic properties. To illustrate the point about word information load, they propose that the findings of Pluymaekers et al. (2010), showing more reduction of the sequence /xh/ in Dutch suffixes when the suffix is divided between /x/ and /h/, can be accounted for by the smaller morphological paradigms for those words. Because the paradigm is smaller, the number of words which are competing is smaller and thus the /xh/ sequence is less important for word identification. However, the definition of word information load used by Hanique and Ernestus is different in some key ways from that used in van Son and Pols, a point which will be discussed further in Chapter 3.

Additionally, recent studies exploring the influence of morphological predictability on the realization of morphemes provide further evidence that morphological structure is

relevant to acoustic realizations (Cohen, 2014, 2015; however see Seyfarth, 2016). These studies control for word-level factors such as word frequency and word bigram predictability, reducing the potential confound with word information load. In Cohen (2014), English 3rd person singular /s/ is found to have longer duration when the relative frequency (referred to here as paradigmatic predictability) of the complex word is higher, a result which is compatible with that of Schuppler et al. (2012). Additionally, for low-frequency verbs, /s/ which is more predictable given the context tends to be shorter relative to the base. In Russian (Cohen, 2015), first person and third person agreement suffixes (*-i, -o*) tend to be more dispersed in the vowel space when the paradigmatic predictability is higher, while their distance on F1 is reduced when contextual predictability is higher. Unlike the studies discussed in Hanique and Ernestus, these studies all control for factors related to word information load, and therefore provide stronger evidence that morphological structure is important to the acoustic realization of morphemes. However, the degree to which this is true, and the nature of the effects, is an area that requires further investigation.

Overall, while there is still some uncertainty related to how morphological structure influences the phonetic realization of morphemes and segments within complex words, there is a significant amount of evidence suggesting that it does play a role, independently of word-level effects. The studies presented in this thesis continue to investigate the role of morphological structure, and provide additional evidence that morphemes do have some degree of independent representation, which influences their realizations. The following section outlines the three studies presented in this thesis, and gives an overview of the predicted results based on extending the predictions of the Rescorla-Wagner model and Message-Oriented Phonology to the learning and production of linguistic material that encodes the abstract message of plurality.

1.5 Research questions and hypotheses

As discussed in Section 1.1, the background presented here on the influences of predictability on linguistic behavior at all levels of linguistic structure (Section 1.2), treating language as a system of message transmission (Section 1.3), and whether morphemes have some degree of independent representations (Section 1.4), leads to questions regarding the nature of the knowledge that language users have about morphemes.

The overarching research question of this thesis is:

RQ 1: What constitutes a language user's knowledge with respect to morphemes?

This question will be addressed by more specifically examining whether the statistical properties of morphemes are tracked in the same way usage patterns are tracked for other levels of linguistic structure. This leads to the more specific formulation of RQ1:

RQ 1a: Do language users have knowledge of the predictability of morphological cues?

RQ 1a will be answered by examining whether linguistic behavior is influenced by the statistical properties (e.g. contextual predictability) of morphological cues, leading to research questions 2 and 3:

RQ 2: How is the learning of linguistic cues which signal the grammatical category of plurality influenced by predictability?
RQ 3: How is the production of linguistic cues which signal the grammatical category of plurality influenced by predictability?

If linguistic behavior is shaped by morphological predictability, on top of word predictability, it will provide evidence in the debate about whether morphemes have representations which are independent to a certain extent, RQ 4:

RQ 4: Do bound morphemes have some degree of representation that is independent of the words to which they are bound?

Finally, if the learning and production of morphological cues is influenced by predictability, which suggests that language users do have knowledge of the statistical properties of morphemes, this raises questions about the nature of that knowledge:

RQ 5: Is the knowledge of statistical properties of morphological cues available at a conscious level?
RQ 6: What is the size of the context used to track the predictability of morphemes?

Based on the background provided in this chapter, the following hypotheses are put forward, corresponding to each research question. Hypothesis 1a is related to overall knowledge of morphemes, while Hypothesis 2 relates specifically to learning, and Hypotheses 3 relates specifically to production. Hypotheses 4-6 are related to the details of this knowledge of morphology, and are addressed through production experiments.

H 1a: Language users do have knowledge of the statistical properties of morphemes.
H 2: Linguistic cues signaling plurality are learned less well when the message they signal is more predictable.

H 3: Linguistic cues signaling plurality are produced with more reduced realizations when the message they signal is more predictable.
H 4: Bound morphemes do have independent representations of some nature.
H 5: Language users have conscious access to knowledge of the statistical properties of morphemes.
H 6: More than one word of context is used to track the predictability of morphemes.

Three studies are carried out to test these hypotheses, using the two frameworks presented in Section 1.3.1. Summaries of each study and the predictions with regard to each research question are presented in the following section.

1.6 Overview of thesis studies and predictions

The three studies presented here address the research questions above using the two frameworks introduced in Section 1.3.1, the Rescorla-Wagner (RW) model and Message-Oriented Phonology (MOP). These frameworks provide empirical ways to test hypotheses about whether the predictability of cues to plurality influences the learning and production of these cues. Study 1 uses the RW model to evaluate the learning of multiple cues to plurality in an Artificial Language Learning experiment. In previous research, the Rescorla-Wagner model has been used to model and make predictions about first and second language learning at the level of the word and sound (e.g. Ellis, 2006a, 2006b; MacWhinney, 1997; Ramscar, Dye, & Klein, 2013; Ramscar et al., 2013). However, it also makes clear predictions about how well different cues to the abstract category of plurality should be learned when these cues have different levels of predictability. Studies 2 and 3 use MOP to evaluate how the duration of New Zealand English plural /s/ varies as a function of the contextual predictability of plurality. MOP outlines principles related to the realization of phonemes and how phonological systems change over time. It draws on research that has shown how the realization of linguistic units including phonemes, words, and syntactic structures is influenced by the predictability of the message given the context. MOP also makes clear predictions about how morphemes should be realized depending on their predictability in context. Hypothesis 1a is addressed in all three studies through evaluation of the learning and production of cues to plurality as a function of their predictability. Hypothesis 2 is addressed in Study 1, while Hypotheses 3 and 4 are addressed in Study 2. Study 3 addresses Hypotheses 5 and 6.

1.6.1 Cue redundancy in learning: Learning multiple cues to plurality in an artificial language

Study 1, presented in Chapter 2, addresses Hypotheses 1a and 2 by investigating how the learning of multiple cues to the morphological category of plurality is influenced by manipulating the redundancy of those cues. In an online Artificial Language Learning experiment, participants are exposed to one of two plural marking systems, each with two cues to plurality, which differ in the relative predictability of the two cues. Study 1 uses the Rescorla-Wagner model (with some modifications) to make predictions about how the frequency with which each cue will be chosen as a plural marker in the test phase will differ across conditions.

In both conditions (*stand-alone* and *co-occurrence*) both Cue A (medial gemination, e.g. ka**n**op ~ ka**nn**op) and Cue B (final gemination, e.g. vapo**l** ~ vapo**ll**) are perfectly predictive of plurality, meaning that any time they occur, the word is plural. In the stand-alone condition Cue A is only present in 2/3 of the trials, while in the co-occurrence condition Cue A is always present. With regard to Cue B, in the co-occurrence condition, Cue B only occurs in conjunction with Cue A, while in the stand-alone condition, Cue B does occur in isolation. The frequency of occurrence of each cue is matched across conditions.

The associative learning literature makes clear predictions about the learning of both Cue A and Cue B. Study 1 is set up as a prototypical blocking paradigm (see Chapter 2), which means that in the co-occurrence condition, learning of Cue B should be blocked by learning of Cue A, because the message of plurality is highly predictable from Cue A alone. Thus, the prediction is that in the test phase, participants in the co-occurrence condition should select answers with final gemination less often than those in the stand-alone condition.

For Cue A, a modification to the RW model (Tassoni, 1995) predicts that because in the stand-alone condition there are some trials where the message is plural but Cue A does not occur, participants in the stand-alone condition should select answers with medial gemination less often than those in the co-occurrence condition.

A final prediction of the RW model is that these two above predictions will hold, regardless of whether training stimuli are presented in two separate blocks or in one fully randomized block.

All of these predictions are borne out, suggesting that language users are sensitive to the redundancy of morphemes in a given system, as determined by the statistical distribution of those morphemes. This result confirms Hypothesis 2, that linguistic cues signaling plurality are learned less well when the message they signal is more predictable through other means. This suggests that language users do track the statistical properties of morphemes, providing evidence in favor of Hypothesis 1a. The observation that the statistical properties of morphological cues influence behavior in an Artificial Language Learning experiment sets the stage for examining the effects of predictability in real language, as the predictability of a given cue varies depending on context.

1.6.2 Cue redundancy in the wild: Gradient phonetic realizations of plural marking in New Zealand English

Study 2, presented in Chapter 3, shifts the focus from learning to production, and looks at plural marking in a natural language. This study addresses Hypotheses 1a, 3, and 4 by examining the realization of plural /s/ in New Zealand English as a function of the morphological predictability of plurality given the context, or the relative amount of information carried by the plural marker in a given context.

In a corpus of New Zealand English, a measure of the morphological predictability of plurality is calculated based on the word preceding the plural word (preceding word plural predictability, PWPP). If the preceding word is often followed by a plural (e.g. *various*), then the PWPP is higher. Likewise, if the preceding word is less often followed by a plural (e.g. *pretty*), PWPP is lower.

According to Message-Oriented Phonology, the prediction is that when PWPP is higher, the duration of plural /s/ will be shorter, and vice versa. In contexts where the probability of successful communication of plurality is already high due to a high PWPP, /s/ duration can be reduced. On the contrary, when PWPP is low, investing resources and producing a longer /s/ is important in order to maintain the high probability of communicating the message of plurality.

This prediction is also borne out, with higher PWPP correlating with shorter /s/ duration. This finding confirms Hypothesis 3, providing further evidence that the statistical properties of morphemes are tracked by language users (Hypothesis 1a), and that language users make use of predictability statistics to transmit the message of

plurality in an effective way. This study also controls for word-level factors such as word frequency and word bigram predictability, providing evidence that morphological predictability is independent of lexical predictability. This evidence supports Hypothesis 4, that morphemes have independent representations of some nature.

1.6.3 How much context matters: Online rating of plural contexts in NZE

Study 3, presented in Chapter 4, expands on Study 2 by both using a different methodology to measure morphological predictability and increasing the size of the context which is used to calculate the predictability of plurality, thereby addressing Hypotheses 5 and 6.

Rather than calculating plural predictability from a corpus, this study uses an online subjective rating task to measure Preceding Context Plural Predictability (PCPP) based on one or five words of preceding context. Preceding contexts of one and five words are extracted from the same corpus used in Study 2, and participants are asked whether a given context is more likely to be followed by a singular or plural noun. These ratings are combined to calculate PCPP scores. By using a subjective rating task rather than a larger corpus, these PCPP scores may capture aspects of the context that would not be well captured by n-gram predictability. They also provide a metric for testing whether language users have conscious access to knowledge about the statistical properties of morphemes (Hypothesis 5).

If the subjective ratings successfully capture morphological predictability, the PCPP scores based on one-word contexts should be highly correlated with the PWPP scores calculated in Study 2. If this is not the case, it will suggest that this methodology is not capturing a language user's knowledge of morphological predictability, perhaps because language users do not have conscious access to these probabilities. In this case, an alternative methodology for calculating subjective PCPP scores should be considered.

If, however, the scores for the one-word contexts are correlated with PWPP, the predictions for Study 3 are similar to those of Study 2. This study continues to use the MOP framework, but with a new definition of message predictability. Thus, /s/ duration should be shorter when PCPP is higher and vice versa. However, this study also introduces the question of how much context is relevant when predicting effects on the realization of an inflectional morpheme, addressing Hypothesis 6. Because information about the message of plurality may be conveyed several words before the plural word (e.g. ***A few*** *incredibly cute dogs****s***), the prediction is that the PCPP based on five-word

contexts will be a better predictor of plural duration than the PCPP based on one-word contexts. This is because PCPP scores based on the five-word context are expected to be a more accurate representation of message predictability.

Hypothesis 5 is supported to a certain extent. PCPP ratings based on one word of context are significantly correlated with PWPP ratings, indicating that language users do have some conscious knowledge of plural predictability. However, the one-word PCPP score is not a significant predictor of /s/ duration, indicating that the PCPP score does not capture true plural predictability successfully. This result indicates that fine-grained knowledge of morphological predictability is not available at a conscious level.

As for Hypothesis 6, the results are not conclusive regarding how large the relevant context is for calculating morphological predictability, as neither the one-word nor the five-word PCPP score is predictive of plural duration. The end of Chapter 4 proposes alternative methodologies which might capture subconscious knowledge of plural predictability, and allow for investigation of the size of the relevant context for calculating morphological predictability.

1.6.4 Conclusions

The final chapter, Chapter 5, presents an overview of the results and discusses the implications of the findings. Hypotheses 2 and 3 are confirmed, with evidence from Studies 1 and 2 demonstrating that morphological cues which are more predictable are learned more poorly and produced with reduced realizations. Both of these results are compatible with the view that language users are using the statistical properties of linguistic units to use language as an effective communication system, and that this extends to the level of the morpheme. This provides evidence in favor of Hypothesis 1, that language users have access to knowledge about the statistical distribution of morphemes. Study 2 also provides evidence in favor of Hypothesis 4, that morphemes have some degree of independent representation. Study 3 provides marginal evidence in favor of Hypothesis 5, and is inconclusive with regard to Hypothesis 6. With regard to Hypothesis 5, the results suggest that language users have some conscious access to morphological predictability, but that this knowledge is not as nuanced as predictability calculated from a corpus. While the Hypothesis 6 was neither confirmed nor disproven, with subjective ratings of plural predictability failing to predict /s/ duration for either context size, several proposals are made which could address this question in a different

way in the future. The Rescorla-Wagner and Message-Oriented Phonology frameworks provide empirical ways to measure the information carried by plural markers both across systems and within a given system, demonstrating that plural marking carrying less information is likely to be learned less well and to be produced with a more reduced realization.

To summarize, this thesis examines the learning and production of bound morphemes as a function of their predictability, using the plural as a case study.[5] The influence of predictability on linguistic behavior related to morphemes is used to address the question of whether language users have knowledge of the statistical properties of morphemes, and thus to inform the question of what constitutes a language user's knowledge of morphemes. This line of enquiry also informs the debate about whether morphemes have independent representations of some nature. Overall, this work contributes to the body of knowledge examining the extent to which language users are aware of the statistical properties of linguistic units, and how those statistical properties influence behavior in a way that is consistent with treating language as a system of message transmission.

[5] While it is likely that the findings extend to other bound morphemes, this is not necessarily true. The plural is among the most productive and most decomposable morphemes, making it a good target for initial examination. However, other morphemes which are less productive, or less decomposable may not show the same effects.

2 LEARNING REDUNDANT CUES TO PLURALITY

2.1 Introduction

As discussed in Chapter 1, this thesis investigates the ways in which the predictability of morphemes influences linguistic behavior related to morphemes. The study presented in this chapter investigates the influence of predictability on the *learning* of morphological cues to plurality, addressing Research Question 2 (see Chapter 1). Hypothesis 2 states that linguistic cues signaling plurality are learned less well when the message they signal is more predictable. The predictability of a morphological message can be measured in several ways, but in this thesis the focus is on how predictable the message is, given the surrounding material. In this chapter, the relevant surrounding material is other cues within the same word, and the usage patterns of those cues across the linguistic system.

Given the way that predictability is measured here, Hypothesis 2 can be restated as:

> **H 2:** Linguistic cues signaling plurality are learned less well when the message of plurality is more predictable, given other cues within the same word.

As a reminder, this is because the predictability of the message signaled by a cue, given other cues within the same word, can be used to calculate the amount of *information* carried by each cue (Shannon, 1948), and thus how important that cue is to communicating the message of plurality. Cues which are more important to successful message transmission should be learned better (see, e.g. Fedzechkina et al., 2017). This

is directly related to the discussion of information and redundancy in Chapter 1, where the communication system is optimized to facilitate successful communication while not expending unnecessary resources. Learning will be measured here by how often words containing a given cue are selected to represent plural images in the test phase. To the extent that this hypothesis is supported, it will suggest that language users track the statistical distribution of morphological cues.

The literature about associative learning, specifically the Rescorla-Wagner (RW) Model (Rescorla & Wagner, 1972, see Section 2.2.1) makes clear predictions about how well multiple cues should be learned, depending on their patterns of co-occurrence. The relevant phenomenon for this study is that of *blocking*, when one cue, which is highly predictive of a certain outcome (or in this case, a certain meaning) on its own, blocks the learning of another cue. Blocking and a related phenomenon, overshadowing, have been used to explain linguistic behavior, such as the difficulty of learning morphological past tense marking in second language acquisition (e.g. Ramscar et al., 2013). According to Dietrich et al. (1995) learners tend to learn to mark tense in a second language using adverbs first. Then, because adverbs are sufficient to convey the message of tense, learning of morphological tense marking is blocked. In artificial language learning experiments, Fedzechkina et al. (2017) find that optional case marking is not used when word order is fixed, showing that, "[if] an existing cue to grammatical function assignment is highly informative, other cues would be redundant and thus could be omitted (24)."

While the RW model has been used to predict linguistic behavior, it has not been used to examine the learning of multiple cues to a given grammatical category, when the predictability of those cues is varied. In this study, it will be used to predict learning of cues to plurality. The associative learning literature does not present learning in terms of effective communication, but the predictions are based on the extent to which a given cue can reduce prediction error about the outcome, which is a measure of how much information is carried by that cue. This means that the predictions of the RW model are aligned with those related to successfully communicating. A more thorough discussion of the RW model is presented in Section 2.1.1.

In order to examine differences in learning morphological cues as a function of the predictability of those cues, it is necessary to have a linguistic system which contains several cues to the same morphological category (e.g. plurality) which differ in their degree of predictability. For example, a language might have one stem-internal

alternation (e.g. doubling of a medial consonant), as well as a suffix, both of which signal plurality. Or, to take an example from English, the past tense might be signaled by a change in the stem vowel (e.g. speed ~ sped) or a suffix consisting of a coronal stop (e.g. hope ~ hoped), or both (e.g. sleep ~ slept). In a given language, these cues might occur always in isolation, always together, or sometimes together but sometimes not. Table 2.1 shows example datasets for three hypothetical languages, drawing from English verb forms. In the table, Cue A is the vowel alternation and Cue B is the presence or absence of a word-final coronal stop.

Table 2.1: Hypothetical Languages with Multiple Cues to Past Tense.[6]

Language 1 (Always separate)			**Language 2 (Sometimes together)**			**Language 3 (Always together)**		
present	past	cue(s)	present	past	cue(s)	present	past	cue(s)
hope	hoped	B	sleep	slept	A, B	sleep	slept	A, B
speed	sped	A	speed	sped	A	keep	kept	A, B
meet	met	A	meet	met	A	leap	leapt	A, B
poke	poked	B	keep	kept	A, B	weep	wept	A, B
feed	fed	A	feed	fed	A	creep	crept	A,B

In Language 1, Cue A and Cue B never co-occur. Neither cue is entirely predictive of the tense of the verb so both carry a relatively large amount of information. However, there is an overall difference in frequency, such that Cue A occurs more often than Cue B. In a language like Language 1, the prediction would be that learners would learn both cues, but will learn Cue A slightly better because it is more frequent. After being exposed to all of these verbs, in a test phase the expectation would be that learners would use Cue A slightly more than Cue B to signal past tense. In Language 2, Cue A can occur in isolation, but Cue B only occurs in conjunction with Cue A. This means that the message signaled by Cue B (past tense) is entirely predictable based on the value of Cue A. For learners of Language 2, the prediction is that they would learn Cue A much better than Cue B, because Cue A is sufficient to distinguish the present and past tense in all cases. Note that although Language 1 and Language 2 have the same number of occurrences of Cue B, the relationship with Cue A is different, leading to the

[6] Past tense is used here rather than plurality because English has clear examples that can be used as two cues to past tense.

prediction that Cue B will not be learned as well in Language 2 as in Language 1. In a comparison between learners of Language 1 and Language 2, the expectation is that learners of Language 2 will use Cue B to signal plurality less than learners of Language 1. Finally, in Language 3, Cue A and Cue B always co-occur, meaning that they are both perfectly predictive of tense. In a language such as this, one prediction would be that learners will learn one cue better than the other, but it would be difficult to predict which cue that would be (assuming equal salience, no pre-existing biases). Learners might alternatively learn the two cues together as a single compound cue.

Another way of looking at the contribution a cue makes to predicting the message, given other cues in the same word, is by using the term *redundancy*. According to Pierce (1980), a cue (or "signal" in Pierce's usage) is redundant if there is more detail in the signal than strictly necessary for communicating the message (plurality) under ideal conditions. In Language 2, for example, Cue A is always sufficient for determining whether the message is present or past. When Cue B is present, it is adding additional detail that isn't necessary, under ideal conditions. In this language, Cue B is redundant. The words in Language 3 also contain redundant material, as there are always two cues to past tense, but it is not clear which of the two cues should be considered the redundant cue. This is because they are both always present, and both sufficient on their own to identify the message of past or present. For the rest of this chapter, the word redundant will be used to refer to cues which signal a message that is highly predictable based on other cues in the same word. Note that a redundant cue, by definition, is carrying less information than a non-redundant cue because the message it signals is predictable from the non-redundant cue.

Making predictions about these simplified languages also allows for predictions to be made about English, which is more complex, but still bears similarities. In English, word-final coronals are the most frequent past tense marker. As in Language 1, the prediction for learners of English would be that the process of adding a final coronal may be learned better than stem alternations. Indeed, evidence of this can be seen in first language acquisition, with children often over-regularizing past tense forms (see Ramscar et al., 2013). However, the English past tense system is not as clear-cut as the sample languages shown above.

In order to investigate differences in the learning of cues based on the predictability of the message they signal given other cues, the ideal situation would be to have two languages, each with multiple cues to a given grammatical category, displaying

different degrees of redundancy. In this way, it would be possible to distinguish effects of frequency from effects of redundancy. In both languages, there would be one more frequent cue and one less frequent cue, but, in the second language, the second cue would be redundant, while in the first language it would not (e.g. Language 2 compared to Language 1 in Table 1). As a reminder, the prediction is that the second cue (e.g. a word-final coronal stop) would be learned less well in Language 2 compared to Language 1 because in Language 2 it is redundant, and therefore not carrying as much information.

However, comparing the learning of multiple cues to plurality across languages raises a number of complications. First, it is difficult to control for other factors, such as phonotactic preferences, which might influence learning. Any two languages will differ on so many levels that attempting to compare the learning of cues to plurality across languages would be very challenging. Second, if real languages were used, the speakers of those languages would enter the experiment with very different language-specific biases which might influence their learning.

In order to avoid these potential complications, a simple artificial language learning (ALL) experiment was chosen as an initial way to examine whether language users are sensitive to usage patterns of morphemes. The experiment is set up as a word-learning game, which will be discussed further in Section 2.2.1. Using artificial languages allows for the comparison of usage patterns across systems while controlling for the effects of language-specific biases. In this ALL design, all participants come from a similar language background, regardless of which system they are exposed to during the course of the experiment. While participants will have some biases related to plural marking and phonotactics, these biases should be consistent across all participants because they all share the same general language background.

This chapter uses the Associative Learning framework, specifically the Rescorla-Wagner (RW, Rescorla & Wagner, 1972) model (with some modifications), to quantify how well two morphological cues to plurality are learned across systems. While there are many theories of cue learning, the RW model is particularly useful for quantifying the learning of cues because it has a clear formula that makes falsifiable predictions. The RW model formalizes how learning occurs as a factor of the information carried by a given cue relative to other cues. A brief summary of associative learning is presented here, before the research questions are stated.

2.1.1 Associative learning

Understanding of associative learning began with the theory of Pavlovian conditioning. In Pavlovian (classical) conditioning (Pavlov, 1927), it is hypothesized that if a certain cue co-occurs with an outcome often enough, that cue will become highly associated with the given outcome. For example, if a mouse repeatedly hears a tone and receives a shock, hearing the tone will cause the mouse to tense in anticipation of a shock.

However, Rescorla and Wagner (1972) examine three phenomena which suggest that mere frequency of occurrence is not sufficient to explain patterns of association in learning, and propose a formal model that accounts for these phenomena. These three phenomena are blocking, overshadowing, and inhibition. All of these phenomena involve learning of multiple cues, and suggest that the associative strength of both cues is relevant in predicting learning patterns. The phenomenon relevant to the present study is blocking,[7] which involves first building a strong association of the outcome with Cue A, followed by trials in which a compound cue, Cue A+B, co-occurs with the outcome. Because the associative strength of Cue A is already so high, B does not gain associative strength. This phenomenon can also occur with intermixed trials of Cue A and Cue A+B, although there will be an initial increase in the strength of Cue B, before it returns to zero.[8]

One of the foundational ideas behind the RW model is that both negative and positive evidence influence learning, and that it is the overall amount of information carried by a given cue in the system that influences its associative strength, not just positive correlations of a cue and an outcome.

In the original RW model, the associative strength of a cue can only change if that cue is present. The strength of Cue A decreases if Cue A is present and the outcome is not

[7] The two other phenomena, overshadowing and inhibition, are not relevant to the present study. Overshadowing occurs when the compound Cue A+B is always presented in conjunction with the outcome, but Cue A is more salient than Cue B. In this case, Cue A ends up with a much higher associative strength compared to Cue B. In inhibition, Cue A is first trained with the outcome, then the compound Cue A+B is presented without the outcome. This results in Cue B having a negative associative strength, which means it inhibits any reaction.

[8] This is because in intermixed trials, the associative strength of Cue A has not yet reached the maximum when the first trial that includes Cue B is encountered.

(e.g. when learning to associate medial gemination with plurality, medial gemination without a plural image). However, there is no predicted effect on the associative strength of Cue A if the outcome occurs without the cue (e.g. a plural image without medial gemination). Kruschke and Blair (2000) point out that phenomena like 'backwards blocking' cannot be accounted for under the RW model. Under RW, a first training phase consisting only in compound cues should result in equal associative strength for the two cues. If the maximum associative strength, the *asymptote*, is reached, then subsequent trials with A alone should have no effect on the strength of A or B. However, empirical evidence demonstrates that backwards blocking paradigms do result in weakened associative strength of Cue B, suggesting a decrease in associative strength for absent cues. There are several proposed modifications to the RW model, which do take into account negative learning for absent cues (e.g. Tassoni, 1995). While these models will not be discussed in detail, the key difference is that the associative strength of a cue can decrease if the target outcome occurs and the cue does not.

Alternatives to the RW model include the contingency model (Cheng & Holyoak, 1995), which successfully accounts for potential changes in associative strength when a cue is absent and the outcome is present. For the present study, the RW model with a modification, as in Tassoni (1995), will be used. However, this is certainly not the only model that would make reasonable predictions. Cheng and Holyoak (1995) provide an analysis of situations in which the RW model and the contingency model make either equivalent or different predictions. For the case of blocking, they are essentially equivalent.

2.1.2 Previous work using associative learning

In the associative learning literature, the following properties have been argued to affect learning of a cue-to-outcome mapping: redundancy (e.g. Ellis, 2006a, 2006b) and availability (e.g. MacWhinney, 1997). The effects of redundancy, as discussed above, are captured by the RW model. Availability, which refers to how often a given cue is present, is captured by the modification to the RW made by Tassoni (1995). A cue which is always available will have greater associative strength.

Evidence from natural (e.g. Ellis, 2006a, 2006b; MacWhinney, 1997; Ramscar et al., 2013) and artificial (e.g. Fedzechkina et al., 2017; Kurumada & Grimm, 2017) language learning suggests that both redundancy and availability do play a role in cue learning.

However, in the above studies, the redundancy comes from semantic information outside the target word. Baayen, Milin, Đurđević, Hendrix, and Marelli (2011) implement the equations from the RW model in a *naive discriminative learning* paradigm, and show that response latency data for complex words can be predicted using cue weights derived from the RW equations, treating every segment in a word as a cue to either a content word meaning or a grammatical meaning. This study investigates whether redundancy based on multiple cues to a grammatical meaning within the same word influences the learning of morphology.

2.1.3 Research Questions

The research questions addressed in this chapter follow from both the overall theme of the thesis, looking at the influence of morphological predictability on linguistic behavior, and the associative learning literature.

RQ 2.1: How is the rate of selection of a cue in the test phase influenced by whether that cue is redundant in training?
RQ 2.2: How is the rate of selection of a cue in the test phase influenced by the availability of that cue in training?
RQ 2.3: To what extent does the order of presentation of stimuli in training (separated or combined training) influence the rate of selection of cues?

These questions are addressed through an ALL word-learning game in which participants must learn cues to plurality. Participants are exposed to one of two artificial languages which display differing degrees of redundancy and availability. In associative learning terminology, the *cues* are the two morphological cues that signal plurality: medial and final gemination. The *outcome* is the message of plurality. *Associative strength* builds throughout the game as participants are exposed to more stimuli. The co-occurrence condition is meant to show *blocking*, so at the end of the game the associative strength of Cue B (final gemination) should be weaker than in the stand-alone condition, resulting in lower rates of selection of Cue B in the co-occurrence condition. The *availability* of Cue A (medial gemination) is different across conditions, meaning that at the end of the game the associative strength of Cue A should be weaker in the stand-alone condition than in the co-occurrence condition, resulting in lower rates of selection of Cue A in the stand-alone condition. Across different training types, or *orders of presentation* of stimuli, the relationship between the learning of cues in the co-occurrence and stand-alone conditions should remain the same, with more selection of Cue B in the stand-alone condition, and more selection of Cue A in the co-occurrence condition. However, when the order of presentation is completely randomized

(combined training, see Section 2.2.4), there should be overall less selection of Cue B than in separated training, across both conditions, with no differences in the rate of selection of Cue A.

The first research question (RQ 2.1) addresses how the rate of selection of a cue is influenced by whether that cue is redundant in training. In the present study, the cues which become associated with the outcome of plurality are medial gemination (Cue A) and final gemination (Cue B) (e.g. kanol ~ ka**nn**o**ll**). The prediction from the RW model regarding RQ 2.1 is that, consistent with Hypothesis 2 (Linguistic cues signaling plurality are learned less well when the message of plurality is more predictable, given other cues within the same word), a cue will be learned less well when it is redundant, given other cues within the word. As stated above, for each trial in which a single cue (e.g. medial gemination, *vannop*) co-occurs with the desired outcome (plurality, +PL), the RW model predicts that the *associative strength* of the cue and the outcome will increase. After a certain number of trials, the associative strength of medial gemination and plurality nears the maximum possible associative strength for a given outcome.[9] In the blocking paradigm, the next stage of training contains two cues together paired with the plural outcome (e.g. medial and final gemination, *kannoll* +PL). Because the associative strength of medial gemination with plurality is already so high, and continues to increase with every trial, there is very little associative strength remaining which can be allocated to final gemination.[10] In this way, medial gemination blocks the association of final gemination with plurality. This prediction is supported by the data.

The second research question (RQ 2.2) addresses how the rate of selection of a cue is influenced by the availability of that cue in training. This question is not related to contextual predictability, but does follow from the associative learning literature, which shows that availability influences learning. Although the classic RW model does not predict any differences in associative strength based on differences in availability, the extension proposed by Tassoni (1995) predicts that a cue which is less available in

[9] If the experiment continued with only trials containing Cue A alone, paired with plural stimuli, then eventually the associative strength of Cue A would asymptote at the maximum associative strength.

[10] If the associative strength of Cue A has reached asymptote when these combined trials begin, then Cue B will not gain any associative strength. This prediction is slightly different when Cue A and Cues A+B trials are intermixed, in combined training.

training will be used less by participants during the test phase. This prediction is supported by the data.

The final research question, RQ 2.3, addresses whether the order of presentation of stimuli during training influences the rate of selection of cues in both conditions. This is addressed by having two types of training. For RQ 2.3, the prediction from the RW model is that the overall relationship between cue learning in the co-occurrence and stand-alone conditions will not be influenced by the order of presentation of stimuli. However, there may be differences in the overall rates of selection of each cue, due to the changes in when the overall associative strength reaches the asymptote. For Cue A, there is no predicted difference, because the associative strength of Cue A should approach the maximum regardless of training type. For Cue B, the prediction from the modified RW model is that, because in combined training, all participants will have some exposure to Cue B before encountering all of the trials that do not contain Cue B, the associative strength of Cue B will decrease for each trial where it is not present. In separated training, the strength of Cue B is zero during all trials that do not contain Cue B, so it is not decreasing for these trials. This will result in a lower final associative strength of Cue B in combined training. This means that overall, there will be less selection of Cue B when training is combined compared to when training is separated. This prediction is not supported by the data.

The remainder of this chapter is organized as follows: Section 2.2 details the methods used in the study, as well as background about using crowd sourcing and game-style experiments for linguistic research; Section 2.3 provides a more detailed discussion of the predictions for the learning of each cue for each combination of training type and condition; Section 2.4 presents the results; Section 2.5 discusses the findings and implications.

2.2 Methods

2.2.1 Game Design

The experiment is implemented as an online word-learning game using a modified adaptive tracking paradigm (Leek, 2001), which was modified for linguistic research by Schumacher, Pierrehumbert, and LaShell (2014). The use of games and crowd-sourcing is discussed further in Section 2.2.1.1. The specific game platform used here has been used previously in word learning studies (e.g. Rácz, Hay, & Pierrehumbert, 2017).

During the game, the participant assumes the role of a bird who is trying to reach its nest, but must fly from rooftop to rooftop before arriving at the nest. At each rooftop, the bird must choose the correct word for a singular or plural picture that is presented. In the training phase, there are two possible answers (see Figure 2.1). If the correct answer is selected, the bird flies to the next roof, and then the next question appears. If the incorrect answer is selected, the bird must go back to the previous roof. Each time the bird moves backwards, the previous question is repeated. Participants are exposed to two different cues to plurality during training.

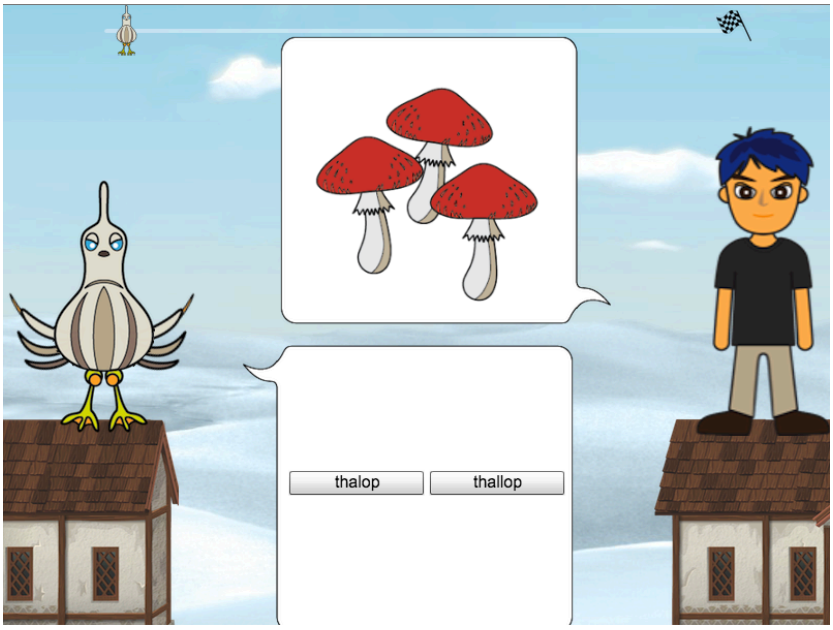

Figure 2.1: Sample screen from the training phase.

In the test phase, the bird has almost made it to the nest, but must answer a series of rapid-fire questions before being allowed to make it home. In this phase, there are four possible answers (see Figure 2.2), including both cues to plurality in various combinations. Regardless of which answer is selected, the bird continues to progress, meaning there is no feedback during the test phase.[11] The following section presents background on the use of crowd sourcing and games in linguistic research.

[11] The participants are informed at the beginning of the experiment that there is no penalty for wrong answers in the test phase. Before the test phase, there is a break in the story and a clear change in the game presentation, with the bird no longer visible.

2.2.1.1 Online word learning

Collecting experimental data online via crowd-sourcing platforms is increasingly common for linguistic research. Crowd-sourcing platforms allow for rapid and low-cost collection of large amounts of data, which might not otherwise be possible. The majority of this research is carried out using Amazon Mechanical Turk (AMT). It has been shown that complex laboratory studies can be successfully replicated with results obtained via AMT, although AMT learners tend to learn less well than laboratory participants (Crump, McDonnell, & Gureckis, 2013).

Specifically, the methodology used in this chapter, called *gamification*, is designed to immerse the subject in the task, and encourage them to pay attention (Von Ahn, 2006). This methodology has been used successfully for linguistic research (e.g. Fedzechkina et al., 2012; Rácz et al., 2017; Schumacher et al., 2014; Toscano, Buxo-Lugó, & Watson, 2015). For example, Rácz et al. find that adult participants learn different contextual meanings of affixes in an artificial language.

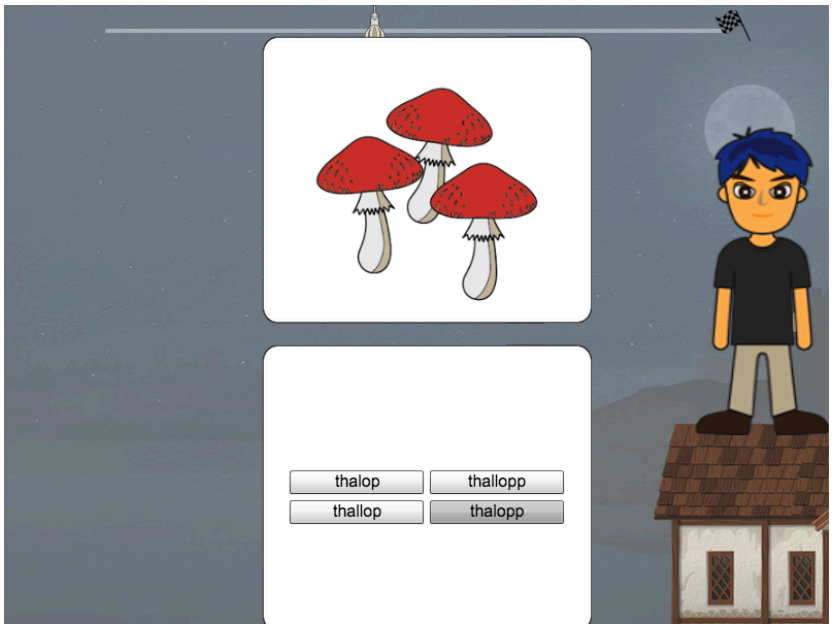

Figure 2.2: Sample screen from the test phase.

2.2.2 Stimuli and cues to plurality

In the game, participants are exposed to two different cues to plurality. The cues are embedded in nonce words which are presented visually using English orthography. Words are combined to form two different artificial language systems with different patterns of usage for plural marking. For each participant, the words which were seen during training and test were randomly selected from a pool of stimuli and randomly

ordered, within a given word type. The co-occurrence patterns of the morphemes in the artificial languages mirror those seen in Languages 1 and 2 from Table 1. A sample of words in each of the two languages is shown in Table 2.2.

Table 2.2: Patterns of Co-occurrence for Multiple Cues to Plurality.

Language 1: Stand-alone condition			**Language 2: Co-occurrence condition**		
singular	**plural**	**cue(s)**	**singular**	**plural**	**cue(s)**
vapol	vapoll	B	vanol	vannoll	A, B
kanop	kannop	A	kanop	kanopp	A
zanop	zannop	A	zanop	zanopp	A
fapol	fapoll	B	fanol	fannoll	A, B
banop	bannop	A	banop	bannop	A

In both systems the two ways of marking plurality are the doubling (visual gemination) of either the medial or final consonant. The majority cue, Cue A, is a geminate medial consonant (e.g. singular: *vanop*; plural: *vannop*). The minority cue, Cue B, is a geminate final consonant (e.g. singular: *bapol*; plural: *bapoll*).[12] For a given participant, only one letter can appear as a medial geminate, and only one letter can appear as a final geminate.[13] The distribution of cues across conditions is explained further in Section 2.2.3. The cues are presented in different combinations depending on the condition (stand-alone or co-occurrence).

[12] The majority and minority cue were not counterbalanced between participants. While a future study should replicate this study with counterbalancing to see if the results hold, it should not bias the results found here. It is possible that one of the cue positions (medial vs. final) may be more salient, which could lead to improved learning. However, the comparison of interest here is not the learning of Cue A compared to the learning of Cue B, but rather comparing the learning of each cue across conditions. If there is a bias that affects Cue B because it is at the end of the word, this bias should apply equally to both conditions.

[13] Until training, where the *p* may occur as a geminate for Type A and Type B in possible responses (see Table 2.3).

Gemination was chosen because it is an observed mechanism for conveying morphological meaning,[14] but is not used as a morphological cue in English. This means that it is an ecologically valid pattern, but participants should not have pre-existing biases about the usage of gemination to signal a morphological category. Additionally, gemination is easily conveyed via orthography, which is necessary because the stimuli in this experiment are presented visually. Although gemination in English orthography does sometimes correspond to changes in pronunciation, these changes are usually in vowel quality (e.g. *put* vs. *putt*). In order to avoid this potential confound, the only vowels used in the experiment are *o* and *a*, which do not display such a change.

2.2.2.1 Singular and plural stimuli

As shown above, the artificial language consists of pairs of written words referring to pairs of images (singular or plural). All singular forms have only singleton consonants, while the presence of geminate consonants in the plural forms varies depending on the pattern. Throughout the discussion, words are referred to by Type, which corresponds to the cues that are contained in the plural form of the word. For Type A words, the plural form contains only Cue A, while in Type B words the plural contains only Cue B. Type A+B words have both medial and final geminates (Cues A and B) in the plural. All stimuli have either a *p*, *l*, or *n* in medial and final position, with a variable initial consonant. Any *l* or *n* is geminate in the plural, while *p* never is. This means that for the majority word type for all participants (Type A), where only words with Cue A alone appear, there is never a final *l* or *n*. In the minority word type for the stand-alone conditions (Type B), there is never a medial *l* or *n*. In the minority word type for the co-occurrence condition (Type A+B), there is never a medial or final *p*. For each participant, wordforms and images were randomly paired to avoid effects of particular word-image pairs.

2.2.2.2 Possible responses

In the training phase, there are two possible responses for any given image. These are always the singular and plural form for a given word type, as shown in Table 2.2. In the

[14] For example in Maltese, gemination is used for several morphological purposes, including distinguishing between the 3rd person and 1st person singular past tense form of certain verbs, e.g. bajjad [ˈbɐjːɐt] ‘he painted’ vs. bajjadt [bɐˈjːɐtː] ‘I painted’ (see Hume, Rose, & Spagnol, 2014).

test phase, there are four possible responses for each image in order to allow for selection of a word consistent with medial gemination, final gemination, both, or neither (see Table 2.3). An example screen from the test phase is shown above in Figure 2.2. A full list of stimuli can be found in Appendix 1.

Table 2.3: Example Response Options for Test Phase.

Word Type	Singular	Plural	Additional options	
A	*banop*	*bannop*	*banopp*	*bannopp*
B	*vapol*	*vapoll*	*vappol*	*vappoll*
A+B	*fanol*	*fannoll*	*fanoll*	*fannol*

2.2.3 Conditions

In order to test the predictions made by the RW model about the correlation of learning with information rather than frequency of occurrence, two main across-subject conditions were run (stand-alone, co-occurrence). In each condition, participants are exposed to two different word types during training. The first word type, which occurs most frequently for all conditions, is word Type A, which contains only Cue A (medial gemination, e.g. *ba<u>n</u>op / ba<u>nn</u>op*). The two across-subject conditions diverge with regard to the second word type which is presented during training. For the stand-alone condition, the second type is Type B, which contains only Cue B (final gemination, e.g. *vapol* / *vapoll*). For the co-occurrence condition, the second type is Type A+B, which contains both Cue A and Cue B (e.g. *fanol* / *fannoll*). In this way, depending on the condition, Cue B is seen either on its own (stand-alone) or in conjunction with Cue A (co-occurrence). The co-occurrence condition follows the pattern of a typical blocking paradigm, as discussed in Rescorla & Wagner (1972). The number of training and test stimuli of each word type for each condition can be seen in Table 2.4 and Table 2.5.

Across conditions, Cue A is the majority cue, meaning it occurs more frequently during training than Cue B. In the *stand-alone* condition, Cue A and Cue B never co-occur during training. Both cues carry a relatively high amount of information in the stand-alone condition because during any given trial either Cue A or Cue B is the sole means of distinguishing a singular from a plural. In the *co-occurrence* condition, all words encountered in training contain Cue A, and thus Cue B only occurs in conjunction with Cue A. Because in the co-occurrence condition Cue B is never the sole means of distinguishing a singular from a plural, it carries less information than in the stand-alone condition. Finally, Cue A has lower availability in the stand-alone condition than in the

co-occurrence condition because there are some trials in which Cue A does not occur, yet the image is plural.

2.2.3.1 Balancing conditions

In order to ensure that the experiment differentiates between effects of frequency and information, balancing of cue exposure across conditions is necessary.

All conditions have one more frequent cue (Cue A) and one less frequent cue (Cue B), but differ in whether the minority cue is presented in conjunction with the majority cue (*co-occurrence* condition), or on its own (*stand-alone* condition). In this way, if similar rates of selection of the minority cue (Cue B) are seen in both conditions, it is the frequency of exposure which determines how well a cue is learned. On the other hand, if the selection rate of Cue B in the co-occurrence condition is lower than in the stand-alone condition, it is likely that reduced selection is due to reduced information carried by the cue, as predicted by both RW model, consistent with treating language as a system of information transmission. This would indicate that language users are sensitive to the statistical properties and co-occurrence patterns of the cues, rather than frequencies alone.

In balancing the co-occurrence and stand-alone conditions, the goal was to ensure that both groups had the same number of training exposures to each cue (medial or final gemination), as well as the same number of total stimuli. However, because in the co-occurrence condition, subjects are exposed to two cues in a single stimulus, it is impossible to balance both total exposures to each cue and total number of stimuli across conditions. This issue led to running two stand-alone conditions. In the first condition (*Stand-alone 1*), the total number of stimuli is the same as that in the co-occurrence condition, as is the number of exposures to the minority cue (Cue B). However, the total number of exposures to the majority cue (Cue A) is different. In the other stand-alone condition (*Stand-alone 2*), the total number of exposures to each cue is the same as in the co-occurrence condition, meaning the total number of stimuli is greater (See Table 2.4 & Table 2.5). In this way, there are two different controls against which the influence of reduced information content can be measured. It is important to

note, however, that due to variation in the amount of time taken to progress through training,[15] the actual number of exposures to each cue varies across participants.

In the test phase, participants see all of the stimuli that they saw during the training phase, as well as unseen stimuli (which were not seen during training) of all three word types (see Table 2.5: 8 unseen stimuli of Type B; 8 of Type A+B; and 16 or 24 of Type A, depending on the condition). This number of exposures ensures that the total number of exposures to Cue B is consistent across all conditions. There is some variation in the total number of exposures to Cue A across conditions, due to the difficulties mentioned above. However, there is no significant difference in behavior between Stand-alone 1 and Stand-alone 2, suggesting that this is not problematic.

Table 2.4: Numbers of Stimuli of Each Type in Each Condition, Training.[16]

	Word Type	Ex.	Stand-alone 1	Stand-alone 2	Co-occurrence
Training	(1) Type A	*banop* / *bannop*	16	24	16
	(2) Type B	*vapol* / *vapoll*	8	8	0
	(3) Type A+B	*fanol* / *fannoll*	0	0	8
	Total medial (Cue A):		16	24	24
	Total final (Cue B):		8	8	8
	Total training stimuli		24	32	24

Table 2.5: Numbers of Stimuli of Each Type in Each Condition, Test.

	Contrast Pattern	Ex.	Stand-alone 1	Stand-alone 2	Co-occurrence
Test	(1) Type A	*banop*/ *bannop*	32	48	32
	(2) Type B	*vapol* / *vapoll*	16	16	8*
	(3) Type A+B	*fanol* / *fannoll*	8*	8*	16
	Total medial (Cue A):		40	56	48
	Total final (Cue B):		24	24	24
	Total test stimuli		56	72	56

* Starred items represent only unseen stimuli, as these are the pattern types that the given condition was not trained on. All other cells in the test phase represent half seen and half unseen items.

[15] When an incorrect answer is given during the training phase, the participant is sent back one question, and has to repeat. This means that certain participants will have more exposures than others.

[16] Numbers of stimuli include words paired with both singular and plural images, which occurred with equal frequency for each word type.

2.2.4 Training type

The combination of three word types (Type A, Type B, Type A+B) and two conditions (stand-alone, co-occurrence) allows for the investigation of whether a given cue to plurality is learned more effectively when it is carrying more information. However, this study also examines whether the effects of word type and condition are influenced by the order of presentation of stimuli.

In order to test whether the influence of information carried by cues is consistent regardless of the order of presentation of stimuli, two different types of training were run.[17] In *separated* training, participants were exposed to two training phases, each of which contained only one word type. For both conditions, in separated training the first training phase contained only Type A words, while the conditions diverged during the second training phase, as detailed above. There was no discernible break between training phases for participants who received separated training. For participants who received *combined* training, all words were fully randomized into one longer training phase (equal to the two phases in separated training combined). A summary of the word types seen in each condition for each type of training can be seen in Table 2.6 and Table 2.7. As a reminder, the prediction is that in the test phase, participants in the co-occurrence condition will select Cue B less than participants in the stand-alone condition, while they will select Cue A more than participants in the stand-alone condition. If these effects are seen across both types of training, this would provide evidence that it is in fact the statistical distribution (both frequency and patterns of co-occurrence) of the cues that causes blocking, rather than the order of presentation, that is important. The combined training more closely approximates exposure to the statistical patterns of language in real life, where words are not learned in isolated phases.

Table 2.6: Training and Test Progression, Separated Training.

Condition	Training 1	Training 2	Test
Stand-alone:	Type A	Type B	Type A; Type B; Type A+B
Co-occurrence:	Type A	Type A+B	

[17] While these two types of training are often referred to in the literature as *blocked* and *intermixed*, the terms *separated* and *combined* are used here to avoid confusion with the phenomenon of *blocking*, which is predicted to occur regardless of training type.

Table 2.7: Training and Test Progression, Combined Training.

Condition	Training	Test
Stand-alone:	Type A; Type B	Type A; Type B; Type A+B
Co-occurrence:	Type A; Type A+B	

2.2.5 Participants

The task was performed by 367 native speakers of American English recruited on Amazon Mechanical Turk (AMT). There were 194 female participants and 173 male, ranging in age from 18 to 68, with a mean of 34. Eight participants were excluded for not completing the task, and seven participants for response patterns that indicated that they were not paying attention to the task,[18] leaving a total of 352 participants whose responses were analyzed. The number of participants in each combination of condition (co-occurrence, stand-alone 1, stand-alone 2) and training type (separated, combined) is presented in Table 2.8.

Table 2.8: Number of Participants in Each Condition.

	Separated training	Combined training
Co-occurrence	57	60
Stand-alone 1	58	63
Stand-alone 2	54	60

2.2.6 Factors

The influence of three factors on the rates of medial and final gemination was tested: CONDITION, TRAINING TYPE, and WORD TYPE. These are summarized in Table 2.9.

Table 2.9: Factors Considered in Analysis.

Factor	Levels / description
CONDITION	Co-occurrence / Stand-alone (1 & 2)
TRAINING TYPE	Separated (two training phases with one word type each) / Combined (one training phase with two word types combined)
WORD TYPE	Type A (only medial gemination), Type B (only final gemination), Type A+B (both)

[18] Participants were excluded for not paying attention if, during the test phase, they selected the same button position for more than half of the trials. Possible responses were distributed randomly across the 4 button positions, so selecting the same button this often indicates they were just clicking.

2.2.6.1 Condition

The two values for CONDITION are *co-occurrence* and *stand-alone.* Although there were two different stand-alone conditions, they were combined during analysis as they did not significantly differ from each other. The predictions are that participants in the co-occurrence condition will select final gemination less than participants in the stand-alone conditions, and that they will select medial gemination more. These predictions follow from the idea that final gemination carries less information in the co-occurrence condition than in the stand-alone conditions, while medial gemination carries more information in the co-occurrence condition compared to the stand-alone conditions.

2.2.6.2 Training type

The two values for TRAINING TYPE are *separated* and *combined*. Participants who received separated training had two training phases (with no discernable break between them) with one word type each. Participants who received combined training had one training phase with two word types combined. The combined training is meant to more closely approximate exposure to linguistic cues in the real world, where language users learn about the statistical distribution of cues as they are exposed to all types of words. Because the statistical distribution of the cues remains the same in separated and combined training, the Rescorla-Wagner model predicts that a blocking effect (less final gemination in the co-occurrence condition) should be manifested across both training types. However, it is possible that the blocking effect will not be as strong in the combined training, if there are not sufficient stimuli to reduce the associative strength of Cue B. The modification to the RW model proposed by Tassoni (1995) predicts that medial gemination should be selected less often by participants in the stand-alone conditions compared to those in the co-occurrence condition for both training types. This outcome is not predicted by the classic RW model, which does not allow for the possibility of reduced associative strength when a cue is absent.

2.2.6.3 Word type

The three values for WORD TYPE, as discussed in Section 2.2.2, are Type A, Type B, and Type A+B. In training, final gemination occurs in Types B and A+B, so it is likely that final gemination will be selected in the test phase more often for these types than for Type A. Likewise, medial gemination occurs in training for Types A and A+B, so it is likely that medial gemination will be selected in the test phase more often for these types than for Type B. There may be interactions of word type and conditions, as

participants in the co-occurrence condition were only exposed to Types A and A+B in training, while those in the stand-alone conditions were exposed to Types A and B.

2.2.7 Analysis

Responses for plural stimuli in the test phase were analyzed with two separate dependent variables. The first was whether the selected answer had a medial geminate and the second whether it had a final geminate.[19] Note that in training, only certain word types displayed medial or final gemination, although all participants were exposed to both medial and final gemination. Responses to singular stimuli were not analyzed due to limited amounts of variation in response selection.

Effects on the selection of medial and final gemination were analyzed using two binomial generalized linear mixed effects models, using the lme4 package in R (Bates, Maechler, Bolker, & Walker, 2015; R Core Team, 2015). The models included the three key factors discussed above, as well as the theoretically motivated interaction of WORD TYPE and CONDITION. A three-way interaction of WORD TYPE, CONDITION, and TRAINING TYPE was also tested, but was not found to significantly improve model fit. In addition to fixed effects, random intercepts by base word and subject were included, as well as a random slope of word type by subject. Factors and interactions were considered significant if the t value was at or above 2, and if an ANOVA comparison between models that included or did not include the given factor or interaction showed a significant ($p < .05$) difference in model fit.

2.3 Predictions

Table 2.10 summarizes the predictions for each combination of TRAINING TYPE, CONDITION, and WORD TYPE. For each training type (separated, combined), the predicted change in associative strength for each cue during training is shown. As discussed in Section 2.1.1, there is a maximum possible associative strength for the outcome of plurality. As that maximum is neared, the associative strengths of cues asymptote, and the changes in associative strength behave differently. As shown in the table, the

[19] While this is underlyingly a multinomial model, a decision was made to use two binomial models both to simplify the interpretation, and because no standard is yet available for mixed multinomial analyses (see Jaeger, Furth, & Hilliard, 2012).

associative strength of Cue B in the stand-alone condition is predicted to be higher than in the co-occurrence condition, regardless of training type. This is the predicted blocking effect, and is expected to correspond to lower usage rates of Cue B in the co-occurrence condition. This is consistent with the prediction that cues which signal a message that is already predictable will be learned less well.

For Cue A, the associative strength is predicted to be higher in the co-occurrence condition than in the stand-alone condition, again regardless of training type. This is because, in the stand-alone condition, trials where Cue A is not present but the outcome of plurality is present will result in a decrease of the associative strength of A (see Tassoni, 1995). On the contrary, in the co-occurrence condition, Cue A is always present when the outcome is plurality. The lower associative strength of Cue A in the stand-alone condition is predicted to correspond to lower usage rates of Cue A in that condition. This is an effect of the lower availability of Cue A.

Finally, the associative strength of Cue B is predicted to be lower when training is combined than when training is separated. In separated training, the associative strength of Cue B remains at zero during the first part of training. After that, Cue B is present for every trial, so its strength is not reduced due to trials where the outcome is present but the cue is not. However, in combined training, each time a trial occurs that contains Cue B, the associative strength of Cue B increases (if the combined strength has not yet reached the maximum). This means that subsequent trials, where the outcome occurs but Cue B does not, result in a decrease in the associative strength of Cue B. This effect is predicted to be equivalent across conditions, so that the relationship between the stand-alone and co-occurrence conditions will still hold. This means that overall, usage rates of Cue B are expected to be lower for combined training.

Table 2.10: Predicted Changes in Associative Strength During Training.

Separated training

	Stand-alone		**Co-occurrence**	
	Words seen	**Predicted change in associative strength**	**Words seen**	**Predicted change in associative strength**
Training 1	Type A	Cue A up	Type A	Cue A up
Training 2	Type B	Cue B up, Cue A down	Type A+B	Cue A up, Cue B up until asymptote Then, Cue B down
Overall prediction	The associative strength of Cue A will be high, but not quite as high as in the co-occurrence condition. The associative strength of Cue B will be low, but higher than in the co-occurrence condition.		The associative strength of Cue A will be at the maximum, at asymptote. The associative strength of Cue B will be lower than in the stand-alone condition.	

Combined training

	Stand-alone		**Co-occurrence**	
	Words seen	**Predicted change in associative strength**	**Words seen**	**Predicted change in associative strength**
Training (combined)	Type A, Type B	Cue A up, Cue B up when seen Cue A down, Cue B down when not seen	Type A, Type A+B	Cue A up, Cue B up when seen, Cue B down when unseen At asymptote, Cue B down
Overall prediction	The associative strength of Cue A will be high, but not quite as high as in the co-occurrence condition. It will be the same as in separated training. The associative strength of Cue B will be lower than in separated training, but higher than in the co-occurrence condition.		The associative strength of Cue A will be at the maximum, at asymptote. The associative strength of Cue B will be lower than in separated training and lower than in the stand-alone condition.	

2.4 Results

2.4.1 Effect of redundancy (Cue B)

Table 2.11 shows the model summary for the final model predicting the selection of final gemination (Cue B).

Table 2.11: Model Summary, Cue B (Final Gemination).

	Estimate	Std. Error	z value	
(Intercept)	-0.187	0.295	-0.631	
condition: stand-alone	1.373	0.316	4.340	***
word type A+B	-0.001	0.180	-0.003	
word type A	-0.963	0.279	-3.446	***
training type: combined	-0.453	0.250	-1.811	.
condition: stand-alone * word type A+B	-0.736	0.208	-3.530	***
condition: stand-alone * word type A	-1.122	0.327	-3.435	***

This model shows an interaction of WORD TYPE and CONDITION, such that for word Types B and A+B, participants in the co-occurrence condition select final gemination less often than those in the stand-alone conditions. Releveling shows that this effect is significant for Types B and A+B, but not for Type A. This interaction is shown in Figure 2.3, and confirms Hypothesis 2 by addressing RQ 2.1. When final gemination carries less information about the message of plurality, participants are less likely to select it. There is no effect of TRAINING TYPE on the rate of selection of Cue B, contrary to the prediction stated above. This may be due to the number of trials.

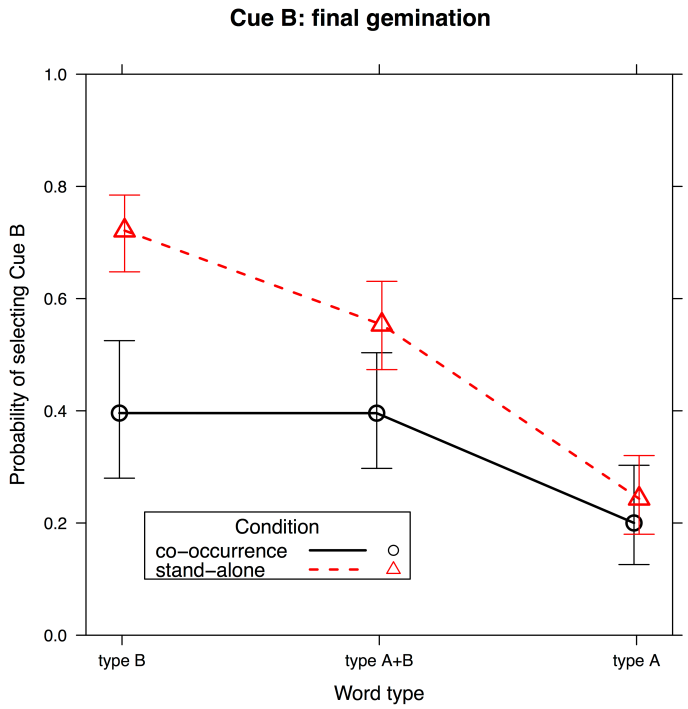

Figure 2.3: Interaction of word type with condition, Cue B

2.4.2 Effect of availability (Cue A)

Table 2.12 shows the model summary for the final model predicting the selection of medial gemination (Cue A).

Table 2.12: Model Summary, Cue A (Medial Gemination).

	Estimate	Std. Error	z value	
(Intercept)	3.296	0.314	10.495	***
condition: stand-alone	-3.513	0.341	-10.312	***
word type A+B	-0.404	0.277	-1.456	
word type A	-0.391	0.319	-1.225	
training type: combined	-0.209	0.187	-1.120	
condition: stand-alone * word type A+B	1.080	0.303	3.560	***
condition: stand-alone * word type A	2.217	0.359	6.178	***

There is again a significant interaction of CONDITION with WORD TYPE (see Figure 2.4). Releveling the model shows that for all word types, medial gemination is selected more by participants in the co-occurrence condition compared to those in the stand-alone condition. This confirms the predictions regarding RQ 2.2 made by the modified

Rescorla-Wagner model (Tassoni 1995).[20] For the co-occurrence condition, there is no significant difference regarding how often medial gemination was selected for the different word types. For the stand-alone conditions, Type A has the most medial gemination, followed by Type A+B and finally Type B. There is also, as with final gemination, no significant effect of TRAINING TYPE.

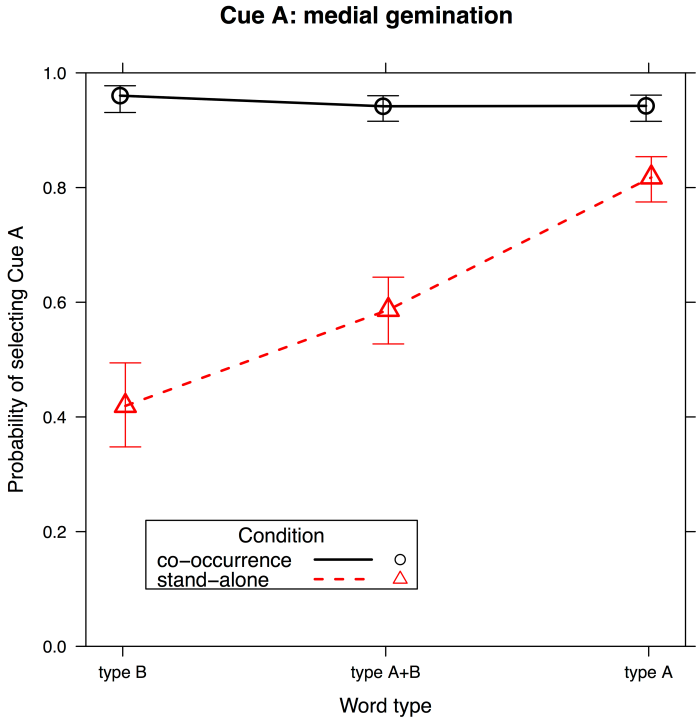

Figure 2.4: Interaction of word type with condition, Cue A.

2.5 Discussion

2.5.1 Cue A and Cue B

The results for final gemination (Cue B) demonstrate that predictability influences the learning of morphological cues. This is shown as a blocking effect, as predicted by the Rescorla-Wagner model (RW; Rescorla & Wagner, 1972). When Cue B only occurs in conjunction with Cue A, and thus the message it signals is more predictable, Cue B is selected less often than when Cue B appears in isolation. This result is consistent across the entire dataset, across both separated and combined training. This result confirms Hypothesis 2:

[20] This is not consistent with the original RW model, as discussed above.

H 2: Linguistic cues signaling plurality are learned less well when they are more predictable.

The observation that this result holds across both separated and combined training suggests that the lower rates of final gemination in the co-occurrence condition are not solely attributable to order of presentation, but are a result of the overall statistical distribution of Cue B compared to Cue A.

The results for medial gemination (Cue A), with lower rates of medial gemination in the stand-alone conditions, are consistent with predictions made by the modified RW model (Tassoni, 1995). Even though medial gemination, when it occurred, was predictive of plurality in the stand-alone conditions, it was not sufficient on its own across all of the data. On the contrary, in the co-occurrence condition, medial gemination was both perfectly predictive of plurality and sufficient on its own to correctly identify the plural in all cases during training.

Both of these findings confirm that in a simple linguistic task, the information carried by a cue and the availability of that Cue in a given system contribute to learning, suggesting that language users are sensitive to the statistical distribution of morphological cues. Final gemination carries less information in the co-occurrence condition compared to the stand-alone conditions, and it is selected less frequently. Likewise, medial gemination is less available in the stand-alone conditions, and it is selected less frequently in the test phase for participants in that condition. Both of these observations support the idea that in learning, behavior is influenced by predictability. This provides support for Hypothesis 1a:

H 1a: Language users have knowledge of the statistical properties of morphemes.

2.5.2 Combined vs. separated training

Participants selected both medial and final gemination at similar rates across combined and separated training. While this was expected for Cue A, Cue B was expected to be used less by participants with combined training. The lack of effect could be due to an insufficient number of trials, or the effect could be so small that it was not able to be captured by rates of selection.

2.5.3 Limitations and future directions

The present study is a stepping stone between simple associative learning tasks and more complex, realistic language tasks involving learning of the distributions of

multiple cues in more linguistically realistic tasks. It is limited in several ways, but provides findings on which future work can build. Future artificial language learning studies could build on the one presented here by varying the reliability of cues, increasing variation in the nonce words, or expanding the multiple cues beyond a single word. Some suggestions follow.

The present study included cues which were 100% reliable. In language, this is rarely the case. A future study with limited reliability of the majority cue could lead to increased learning of the minority cue, even in the co-occurrence condition.

Another limitation was the limited variation in stimuli. The stimuli were chosen to try to reduce the possibility of participants pronouncing stimuli differently in their heads, but this led to a very narrow range of variation. A future study could increase variability in the stimuli and see if similar effects are found.

Finally, the current stimuli design with two cues to plurality in the same word is not typologically common. A future study might spread the cues over a wider range of linguistic material. For example, an adjective and a noun which are both either marked or unmarked for plurality might provide more typologically realistic stimuli while still exploring the same ideas.

2.5.4 Conclusions

While the present study has some limitations, it provides evidence that language users are sensitive to the statistical patterns of use of morphological cues. This sensitivity is evidenced by the systematic variation in rates of medial and final gemination in the test phase of this experiment, with more predictable cues learned less well. This supports Hypotheses 1a and 2. This study also adds to the literature demonstrating blocking effects in morphological learning consistent with the predictions made by the Rescorla-Wagner model. It builds on work by Ramscar et al. (2013) and others, and lays the groundwork for future studies examining learning of multiple cues in a variety of ways which more closely approximate daily language use. Additionally, this study provides evidence that language users are sensitive to the statistical patterns of use of morphological cues. This sensitivity is evidenced by the systematic variation in rates of medial and final gemination in the test phase of this experiment.

The findings presented in this chapter demonstrate from one angle that language users track the statistical properties of morphemes, and that these statistics influence behavior.

The next two chapters approach the question from a different angle, using a corpus of New Zealand English to examine the degree to which plural /s/ duration is conditioned by the predictability of the plural in context. While the present study uses two cues within a single word to manipulate the predictability of morphemes, the studies in chapters 3 and 4 use one cue within the word (the plural /s/), and one external cue (predictability in context) to examine the trade-off of message predictability and signal specificity.

3 CUE REDUNDANCY IN THE WILD: PLURAL MARKING IN NEW ZEALAND ENGLISH

3.1 Introduction

Chapter 2 establishes that in an artificial language learning setting, learning of plural marking is influenced by the predictability of the message of plurality, given other potential cues within the word. Using the Rescorla-Wagner model (Rescorla & Wagner, 1972) and the ideas of associative learning, Chapter 2 shows that when predictability is high due to consistent presence of the majority cue, the minority cue is used less frequently in the test phase. However, when predictability is lower, the minority cue is used more frequently. Likewise, when the majority cue is not always present in training, it is used less frequently compared to when it is always present. These findings suggest that the biases related to treating language as a system of information transfer, namely maintaining a high probability of successful communication while keeping resource costs low, apply to language learning, specifically at the level of the morpheme. While the study in Chapter 2 shows that predictability influences the learning of bound morphemes, it was an artificial setting, with no noise.

The present study builds on Chapter 2 by both examining a real language phenomenon and looking at synchronic variation rather than variation in learning. Additionally, the domain of morphological predictability is expanded beyond the target word to the surrounding context. Using the Message-Oriented Phonology (MOP) framework to

examine effects on the duration of the New Zealand English (NZE) plural /s/ morpheme, this study examines the influence of morphological predictability on the realization of bound morphemes. As a single-segment morpheme,[21] /s/ is the only segment in a regular plural word which is explicitly conveying the idea of plurality.[22] This means that if there are effects based on morphological predictability, they are expected to be manifested in this segment. Indeed, this study finds an influence of morphological predictability on the realization of /s/, with higher morphological predictability correlating with shorter /s/ duration.

Morphological predictability is operationalized in this chapter using the measure Preceding Word Plural Predictability (PWPP), which measures how likely a plural word is to occur, given the preceding word. This measure will be further explained in Section 3.3.3. The prediction is that plural /s/ is realized with shorter duration when a plural noun is more likely to occur given the context. There has been some debate in the previous literature regarding whether effects on morphemes are in fact attributable to factors related to the whole word (see e.g. Hanique & Ernestus, 2012; Hay, 2004; Plag et al., 2017; Schuppler et al., 2012). In order to differentiate from possible effects on the whole word, this study also examines whether plural /s/ duration varies independently of base duration. If the hypothesis that plural /s/ is shorter when it is more predictable is upheld, it may suggest that the principles of effective information transfer are operating at the level of the morpheme.[23] This chapter specifically addresses two of the Research Questions presented in Chapter 1:

RQ 3: How is the production of linguistic cues which signal the grammatical category of plurality influenced by predictability?
RQ 4: Do bound morphemes have some degree of representation that is independent of the words to which they are bound?

[21] Note that [ɪz] realizations of the plural morpheme are excluded in this study.

[22] Some studies have shown that participants are sensitive to differences in the acoustic properties of the base in complex vs. simple words (e.g. *bake* vs. *baker*, Kemps et al., 2005a; 2005b), so it is possible that there are other cues to plurality earlier in the word. However, /s/ is the only phonological cue to plurality in the word.

[23] As discussed in Chapter 1, there are other potential explanations for probabilistic reduction (see Jaeger & Buz, 2017). In this thesis, probabilistic reduction is discussed in terms of facilitating information transmission.

In addressing these questions, this chapter also addresses the overall Research Question:

RQ 1a: Do language users have knowledge of the predictability of morphological cues?

In this chapter, these broad research questions are addressed through the more specific research questions presented here:

RQ 3.1: Does New Zealand English (NZE) plural /s/ duration vary systematically based on how predictable the plural message is, given the context?
RQ 3.2: Does NZE plural /s/ duration vary independently of base duration?

These research questions will be addressed using the Message-Oriented Phonology framework (MOP), developed by Hall, Hume, Jaeger, & Wedel (submitted). MOP builds on previous work, continuing to formalize the relationship between the probability of accurately transmitting a message, the predictability of the message in context, and the degree to which the signal used can uniquely identify the intended message (signal specificity). MOP places the emphasis on conveying messages, rather than individual segments (although see also, e.g. Flemming, 2004, 2010; Kirov & Wilson, 2013). Drawing on an abundance of work that has independently shown the influence of factors such as frequency and predictability in language, MOP brings all of these pieces of evidence together to create a theory of phonology in which phonological systems emerge in the service of effectively using a signal to communicate a message. Further information about MOP and how it is used in this study is provided in Section 3.1.1.

In addition to examining morphological predictability, the study presented in this chapter provides further evidence in support of factors which have been previously shown to have effects on segment duration. These factors are related to predictability (e.g. frequency, bigram word predictability), extralinguistic factors (e.g. speech rate, base duration), and phonological context (e.g. preceding and following segmental context, similarity avoidance effects, and syllable structure). By examining the influence of structural factors, extralinguistic factors, and those related to predictability, this study demonstrates that the pressures exerted by multiple domains can have an influence on phonetic realizations simultaneously. For example, in the domain of predictability, pressures to communicate effectively might result in longer /s/ duration when the message of plurality is less predictable, while simultaneously in the domain of phonology, the pressure of similarity avoidance might result in shorter /s/ duration when there is a similar consonant in the onset.

3.1.1 Message-Oriented Phonology

As discussed in Chapter 1, the MOP framework formalizes the idea that the probability of accurately transmitting a message is a function of the message's predictability and its signal specificity (Hall et al., submitted), though these interacting forces must also be balanced against the consideration of resource costs, which will be further addressed below. The formula, an application of Bayes' Rule, shows the relationship of accurate message transmission, message predictability, and signal specificity is as follows (from Hall et al., submitted):

$$p(M|S, ctxt) = \frac{p(M|ctxt) * p(S|M, ctxt)}{\sum p(M_i|ctxt) * p(S|M_i, ctxt)}$$

For the purposes of this study, *M* is the message of plurality, and the semantically competing message is that of singularity. *S* is the acoustic signal corresponding to the plural /s/, and *ctxt* (context) is the word preceding the target plural word. While according to MOP, the 'message' can be any meaning-bearing unit, most work on phonetic reduction has assumed the 'message' to be a word. This study, in examining a morpheme, extends the concept 'message' to apply to an abstract grammatical category.

On the left side of the equation is the probability of the message *M* given the signal *S* and the context *ctxt*, that is the probability of accurate message transmission. In a communicative context, one can imagine that there is some minimum threshold value below which *p(M|S, ctxt)* should not drop, such that the factors on the right side of the equation should be balanced to avoid going below that minimum. Depending on how important it is to successfully transmit a given message, this number might be higher or lower for different messages, or in different contexts. For the purposes of this study, the assumption is that transmitting the message of plurality is always approximately equally important, so the minimum *p(M|S, ctxt)* is constant.

The right side of the equation consists of two main components: message predictability and signal specificity. Message predictability is the probability of the intended message *M* given the context, while signal specificity is the probability of the signal S given the message and the context. Both of these are normalized by the sum over all possible messages of the probability of the signal given message M_i in the given context, multiplied by the probability of M_i given the context.

Message predictability varies depending on the context. In this study, the context is operationalized as the word preceding the plural, and so *message predictability* is

measured using Preceding Word Plural Predictability (PWPP). In terms of information, if the plural is more predictable given the context (higher PWPP), it is carrying less information.

Signal specificity is a measure of how specific a given acoustic signal is to the intended message, as opposed to all competing messages, and will be measured in this study using the duration of the plural /s/. The only two possible messages considered here are plurality and singularity. Longer /s/ duration makes the signal more specific to the message of plurality by differentiating the signal to a greater degree from the absence of /s/, which would indicate a singular. It is here that the denominator of the equation above becomes important. With no denominator, the equation would read:

$$p(M|S, ctxt) = p(M|ctxt) * p(S|M, ctxt)$$

In this modified equation, if the left side (probability of accurate message transmission) is held at a constant value, then when *p(M|ctxt)* is lowest, or the message is unpredictable given the context, *p(S|M,ctxt)* is highest. However, with no normalization, *p(S|M,ctxt)* is highest for the most frequent realization of the message of plurality (which is most likely an /s/ of medium duration), so the modified equation predicts that this realization will be used in unpredictable contexts. This is problematic because the most frequent realization and the intuitively expected realization in an unpredictable context are not the same; the intuitively expected realization of the message in an unpredictable context is a hyperarticulated signal (a longer /s/). Adding the denominator changes the prediction, by selecting not simply the signal with /s/ duration that is most likely to correspond to the plural message, but one that is likely to be used for the plural message and unlikely to be used for the singular message. Because, within the confines of typically observed /s/ durations, longer /s/ duration makes the signal even more unlikely to correspond to the singular, longer /s/ is predicted in contexts where the message of plurality is not highly predictable.

The two main pressures acting on this system are that of keeping the probability of accurate message transmission high and that of keeping resource cost as low as possible. Resource cost is not represented in the formula, but can be represented by any number of things including time, energy, and processing costs. For the purposes of this study, differences in resource cost will be measured by differences in the duration of plural /s/. The assumption is that more time and energy are invested in producing a

longer /s/. This means that increasing signal specificity also results in increased resource cost.

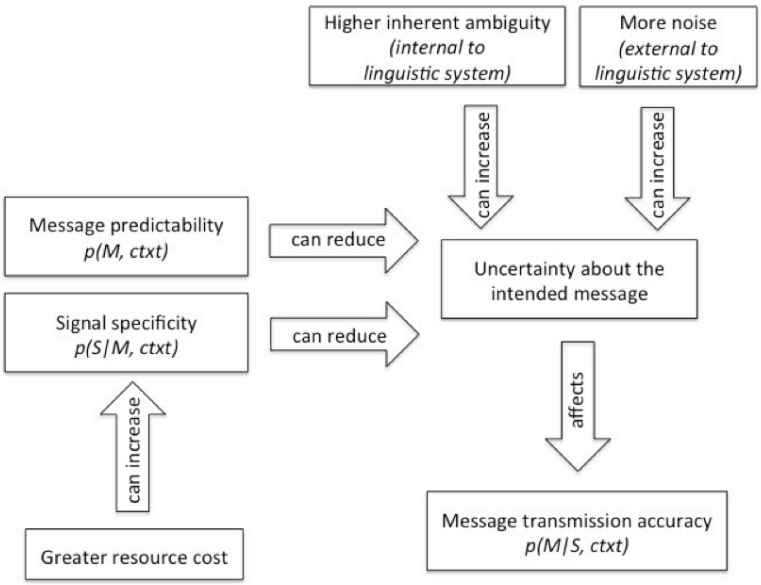

Figure 3.1: The trade-off between message transmission accuracy and resource cost, reproduced with permission from Hall et al. (submitted)

Figure 3.1 shows the influence of both message predictability and signal specificity on uncertainty about the intended message, which is in turn related to message transmission accuracy. It also shows the relationship between resource cost and signal specificity. The two factors at the top of Figure 3.1, inherent ambiguity and noise, are assumed to be constant for the purposes of this study.

Given the two pressures of keeping the probability of accurate message transmission high and keeping resource cost low, the ideal system will balance message predictability and signal specificity to maintain a given probability of successful message transmission. Signal specificity will only be increased (thus increasing resource cost) if message predictability is low. When message predictability is higher, signal specificity can be lower, lowering resource cost. In the present study, the ideal probability of successful message transmission is held constant and message predictability varies based on PWPP. The trade-off of message predictability and signal specificity is expected to be reflected by shorter /s/ duration when the plural is more predictable (higher PWPP), and longer /s/ duration when the plural is less predictable (RQ 3.1). To the extent that this effect is related to the predictability of the morpheme, rather than the

predictability of the word as a whole, the effects on /s/ duration should be independent of variation in base duration (RQ 3.2).

The remainder of this chapter is organized as follows: Section 3.2 provides further background on predictability and morphological structure; Section 3.3 outlines the corpus used for this study, the hypothesized novel factors and control factors known to influence /s/ duration, and the analysis; Section 3.4 presents the results. Discussion of the results in Section 3.5 situates them in the context of predictability, morphology, and phonology.

3.2 Background

3.2.1 Contextual predictability at other levels of linguistic representation

The effect of predictability on the enhancement and reduction of linguistic forms is well documented and has been shown at various levels of linguistic structure. This work is summarized in Chapter 1. While there are studies which address morphological predictability, both paradigmatically (e.g. Baayen, Dijkstra, & Schreuder, 1997; Cohen, 2014; Erker, 2010; Hanique, Schuppler, & Ernestus, 2010; Hay, 2001, 2004; Hundley, 1987; Kuperman, Pluymaekers, Ernestus, & Baayen, 2007; Poplack, 1980; Schuppler et al., 2012; Torreira & Ernestus, 2012) and syntagmatically (Cohen, 2014, 2015; Seyfarth, 2016), a clear picture of the effect of predictability on morphemes has not yet emerged, given differing methodologies and assumptions in past studies, as discussed in Section 3.2.2 below.

3.2.2 Morphological predictability and phonetic realizations

3.2.2.1 Defining morphological predictability

As discussed in Chapter 1, previous work examining the realization of bound morphemes through the lens of predictability, although limited, has looked at both whole word predictability (e.g. Pluymaekers et al., 2005a) and morphological predictability. The influence of *morphological* predictability is evidenced through effects on optional contraction (Bybee & Scheibman, 1999; Frank & Jaeger, 2008), optional case marking (Kurumada & Jaeger, 2015; Norcliffe & Jaeger, 2016), and reduction of both stems and bound morphemes (Cohen, 2014, 2015; Erker, 2010; Hay, 2004; Hundley, 1987; Poplack, 1980; Torreira & Ernestus, 2012).

In order to examine the morphological predictability of bound morphemes, either *paradigmatic* or *syntagmatic* morphological predictability can be used. Paradigmatic morphological predictability measures how predictable a bound morpheme is in a certain word, given the paradigm of words with the same base, while syntagmatic morphological predictability measures how predictable a bound morpheme is given the context in which it occurs. In the terminology of the MOP framework, changes in both syntagmatic and paradigmatic predictability result in changes in message predictability, but in different ways. For syntagmatic predictability, the context surrounding the morphologically complex word makes the given morpheme (e.g. plural /s/) more or less predictable as a message. On the other hand, for paradigmatic predictability, it is the relative probability of /s/ occurring after the base word, compared to any other morpheme, which influences message predictability. In this study we focus on syntagmatic predictability though see, e.g. Cohen (2014, 2015), Hay (2004), Pluymaekers, Ernestus, & Baayen (2005b) for further discussion of paradigmatic predictability.

In order to calculate syntagmatic morphological predictability, methods used for calculating bigram word predictability can be extended (see Section 3.3.4.1.2). Bigram word predictability is calculated by taking the number of times that a given word w_i occurs after word w_{i-1}, then dividing by the total occurrences of word w_{i-1}. Similarly, the conditional probability of a bound morpheme m can be calculated given the word preceding or following the complex word ($P(m_i|w_{i-1})$, $P(m_i|w_{i+1})$). In the following equation, m_i represents any word i containing the morpheme m, so the numerator is the frequency with which the word before the complex word occurs in combination with any complex word containing that morpheme.

$$P(m|w_{i-1}) = \frac{freq(w_{i-1} + m_i)}{freq(w_{i-1})}$$

For inflectional morphemes such as English plural /s/, this measure allows us to calculate how probable a plural noun is to follow or precede a given word; that is, the probability of the plural morpheme. For example, in the phrase *those cats*, the plural morpheme is highly likely to occur after a plural determiner, whereas in *the cats*, either a singular or a plural is possible, so the probability of the plural morpheme is lower *(P(-s | those) > P(-s | the)*.

The following section discusses other studies which have investigated syntagmatic morphological predictability, although many of them measure morphological predictability in different ways.

3.2.2.2 Previous work involving syntagmatic morphological predictability

Work involving syntagmatic morphological predictability has been mainly in Spanish and English. A variety of research on Spanish word-final /s/ has been undertaken, as well as work on English third-person singular /s/.

There has been a significant amount of work exploring morphological /s/ reduction in various dialects of Spanish,[24] much of which is directly related to syntagmatic morphological predictability and phonetic reduction.

These studies operationalize morphological predictability as the extent to which the /s/ suffix is morphosyntactically redundant. This is determined by looking at whether the target noun is preceded by either another word marked for plurality (e.g. *las*, a plural article), or some other word that indicates plurality (e.g. *cuatro*, 'four'). Poplack (1980) and Hundley (1987) find that in Puerto Rican and Peruvian Spanish, respectively, /s/ is more likely to be deleted when redundant. Erker (2010) examines the gradient realization of Spanish /s/, exploring subphonemic detail, and finds that morphosyntactically redundant /s/ tends to have shorter /s/ duration. Torreira and Ernestus (2012) examine intervocalic /s/, using other gradient measures of /s/ reduction: the duration of a low-band intensity dip between the two vowels, voicing of the /s/, and the difference in high-band intensity between the /s/ and the surrounding vowels. They find no significant differences in either intensity measure, but a marginal difference in voicing, with redundant affixes realized as voiced less often than nonredundant affixes. This provides some evidence that /s/ phonemes which are more predictable given their context are more likely to be reduced.

[24] This work includes research regarding morphological predictability, but also research examining differences between morphemic and non-morphemic /s/ (e.g. Hundley, 1987; Poplack, 1980). This is related to work in English that has compared morphemic and non-morphemic /t/ and /d/, as well as /s/ and [z] (Plag et al., 2017; Seyfarth, 2016; Walsh & Parker, 1983). These studies have found conflicting results in terms of whether morphemic or non-morphemic segments tend to be longer.

In addition to the research on Spanish /s/, there has been some research examining syntagmatic morphological predictability in English. Cohen (2014) and Seyfarth (2016) evaluate the potential influence of paradigmatic and syntagmatic predictability on the duration of the third person singular agreement suffix on English verbs. Most relevant to the current study, Cohen operationalizes syntagmatic predictability as the probability of agreement attraction on verbs in sentences with complex subjects (see Bock & Miller, 1991). Agreement attraction is a phenomenon in which the main verb in a sentence with a complex subject agrees with a nearby noun phrase, rather than the head noun. For example, a speaker might say "The description of the pictures are beautiful," using the third person plural form of the verb even though the head noun is singular. Depending on the syntactic structure (Bock & Cutting, 1992; Franck, Vigliocco, & Nicol, 2002), length (Bock & Cutting, 1992; Bock & Miller, 1991), and semantic properties (Humphreys & Bock, 2005; Solomon & Pearlmutter, 2004; Thornton & MacDonald, 2003) of the complex subject, the likelihood of agreement attraction varies (for a cross-linguistic review, see Jaeger & Norcliffe, 2009). Cohen records a range of sentences using preambles which have been previously shown to vary in their likelihood of eliciting agreement attraction (e.g. "The pizza with the missing slices looks unappetizing in the morning basket"). The sentences are read aloud by participants after they have read and processed the meaning of the sentence. The likelihood of agreement attraction occurring in the given sentence is then used as a measure of contextual morphological predictability to predict /s/ duration on the verb. The sentences where a plural verb form is most likely are those where the head noun is a collective plural such as "jury" or "class".

This method of calculating morphological predictability differs significantly from the way that contextual predictability is usually calculated for other levels of linguistic structure, which could contribute to the absence of an effect on suffix duration (see below). This method does measure the contextual predictability of the /s/, but does not capture differences in the contextual predictability of the meaning. In the present study, the measure of morphological predictability that is used (PWPP) corresponds more closely to traditional conceptions of contextual predictability, and focuses on the predictability of the meaning. Seyfarth (2016) uses a measure of contextual morphological predictability based on a cloze norming task, in which participants were given the beginning of a dialogue, ending directly before the target word, and asked to fill in any word. Contextual morphological predictability is then calculated based on

how many participants used any 3sg verb in a given context. While this does capture morphological predictability, focusing on meaning, the set of contexts used was designed to test differences between morphological and non-morphological word-final segments, so may not provide a wide range of values for morphological predictability.

In terms of syntagmatic predictability, Cohen finds no effect on the absolute duration of the suffix, but does find that for low frequency verbs, increased syntagmatic predictability correlates with relatively shorter suffixes. For high frequency verbs, this effect is either not present, or in the reverse direction. Although these results suggest that contextual morphological predictability correlates with reduction, the sentences used in Cohen's study were chosen specifically to create a continuum of syntagmatic predictability, and were recorded in an experimental setting. While this allows for greater control in terms of experiment design, it may be the case that because subjects were reading as opposed to speaking spontaneously, they behaved differently than they would have had they been speaking naturally. Seyfarth (2016) finds no significant influence of syntagmatic predictability on 3sg /s/ duration, but likewise uses experimental data and a set of contexts not designed to address this question. Additionally, the 3sg suffix is highly constrained by grammatical context. While this suffix does occur frequently, the contexts in which it is possible for speakers to use either the 3sg or 3pl verb form are very specific, and still prescriptively require one or the other. The trade-off between message predictability and signal specificity is most important when there is some question as to the intended message. It is for this reason that ambiguous contexts are essential. In clearly constrained grammatical contexts, for example those used by Cohen, there is very little contribution that the suffix itself can make to add information; the potential change in uncertainty is low. However, in ambiguous contexts, there is more room for variation, and a higher potential change in uncertainty. For this reason, examining an inflectional morpheme which is less constrained by context may provide further insights. For English plural /s/, for example, there are many contexts in which plural and singular nouns are both licit. This allows for many situations in which the context is ambiguous regarding plurality, and thus the role of predictability can be highlighted.

The present study builds on this previous research by using spontaneous speech recordings to study variation in the realization of English plural /s/. It also calculates contextual morphological predictability of plural meaning rather than the segment itself, in a way that is directly connected to the typical formulation of contextual word

predictability, using corpus statistics. This measure is likely to more accurately capture the intuitions of language users with regard to what contexts plurals are likely to occur in, and can be calculated for any context in the corpus. In contrast, the measure used by Cohen (2014) can only be calculated for a select set of sentences which have specific properties, and is based on previous experimental results. Additionally, unlike 3sg /s/, there are many contexts where either a singular or plural noun could be correct (e.g. *look at the cat/cats*). These ambiguous contexts allow for the investigation of the influence of the predictability of plurality given the context on the duration of plural /s/. While some studies have investigated plural /s/ in Spanish, these have all measured predictability categorically, and only a few have looked at gradient phonetic properties of the /s/. Spanish is also quite different from English, in that plural marking is not required on adjectives in English, although determiners and verbs may provide morphosyntactic redundancy. By using over 5000 tokens of plural /s/ from spontaneous speech and investigating a different inflectional morpheme, this study aims to provide further evidence that the phonetic realization of short morphemes is affected by contextual morphological predictability.

3.2.2.3 Morphological predictability vs. *word information load*

As noted in Chapter 1, there is an on-going debate about whether or not morphemes have some sort of representation that is independent of the whole word in which they occur. This is highly relevant to this study, because if morphemes do not have independent representations of some sort, language users will have difficulty tracking their predictability in context. If morphemes are only processed as a part of the entire word, a measure of morphological predictability that does not take the base word into account should have no influence on the duration of morphemes. Most relevant to this study is the acoustic evidence suggesting that morphological structure influences the phonetic realizations of morphemes and segments in complex words. One of the main arguments against the idea that morphological structure influences the phonetic realization of morphemes is raised by Hanique and Ernestus (2012), who suggest that effects which appear to be due to morphological predictability can actually be explained by what they call *word information load* (van Son and Pols, 2003). Word information load, as formulated by van Son and Pols (2003), measures the contribution a given segment makes to identifying the word as a whole, either with or without taking into account the 'context distinctiveness' of the word. However, Hanique and Ernestus use a

slightly different definition of word information load when arguing that morphological predictability does not play a role.

In van Son and Pols, the measure of 'context distinctiveness' is essentially the average predictability of the target word calculated across a corpus, given up to 5 words preceding and following the target word. This measure is constant for any given word, calculated over a given corpus. Hanique and Ernestus, on the other hand (reanalyzing data from Torreira & Ernestus, 2010), include morphosyntactic information in their conceptualization of word information load, claiming that the contribution a morpheme makes to identifying the word is influenced by whether there is information in the surrounding context that indicates plurality (e.g. a plural article or a number). Contrary to the measure of word information load used by van Son and Pols, this does vary by context, so is not constant for a given word. Additionally, the presence or absence of material that indicates plurality is only partially related to identifying the word as a whole, and is intricately related to identifying the morpheme.

If Hanique and Ernestus are correct, and it is solely identification of the word as a whole that matters, rather than the morpheme /s/, a measure of word n-gram predictability should be able to account for the variation in /s/ realization. Additional measures of morphological predictability should then not add further explanatory power. While this is not explicitly tested in Hanique and Ernestus, it would be interesting to see whether word predictability explains all of the variation in Torreira and Ernestus.

The present study seeks to differentiate the two causal factors of morphological predictability and word information load (or the contribution a segment makes to identifying the whole word) by examining the effect of morphological predictability on a bound morpheme, the plural marker /s/, which occurs repeatedly in the same words throughout the dataset in contexts which vary in both bigram word predictability and morphological predictability. If contextual morphological predictability is shown to contribute to systematic phonetic variability of the morpheme within a given word beyond the contribution of bigram word predictability, the results would indicate that properties of bound morphemes affect their realization independent of their base words. This in turn would provide further evidence that morphemes are psychologically real linguistic units.

3.3 Methods

3.3.1 Plural /s/

In testing whether paradigmatic morphological predictability has an effect on gradient realizations of bound morphemes, NZE plural /s/ was selected for several reasons: it is an inflectional morpheme; occurs frequently in naturally occurring speech; is not entirely constrained by grammatical context; consists of only one segment; and has been shown, on average, to have longer duration that other /s/ morphemes (Plag et al., 2017; Zimmermann, 2016).

One advantage of using an inflectional morpheme as opposed to a derivational morpheme is that calculating the syntagmatic predictability (contextual predictability based on words external to the base word) of the former is more tractable. While derivational morphemes tend to be more restricted regarding which base words they can attach to, inflectional morphemes are much more productive. This means that for inflectional morphemes, it is easier to abstract away from word-specific predictability in order to focus on predictability based on the base word-external context. Since all regular verbs, for example, can take a past tense morpheme, it is possible to calculate how likely a given word *w* is to be followed by a past tense verb (this could be called preceding word past tense predictability, for example). This can be calculated over either all instances of *w*, or over only instances of *w* when followed by a verb of any type. On the contrary, if we want to see how likely *w* is to be followed by a verb with an *un-* prefix, only the likelihood over all instances of *w* can be used, since calculating over all instances of *w* followed by a verb is not actually a measure of the likelihood of *un-,* which cannot attach to all verbs. This is because not all verbs can have *un-* prefixes. Additionally, in order to treat the affix consistently as a meaning-bearing unit, the affix must have a transparent and uniform meaning across words. While this is usually true for inflectional affixes, derivational affixes are variable, and do not always have transparent meaning (e.g. *tasteless* vs. *listless;* Hay & Baayen, 2002).

Another reason why plural /s/ is the object of study is that it occurs frequently in naturally occurring speech. Given the focus of this thesis on the communicative biases that influence learning and speech production, naturally occurring speech is an appropriate source of data. In natural speech data, speakers are almost certainly trying to communicate messages to other language users, rather than performing an experimental

task. Further, because plural /s/ occurs frequently in spontaneous speech, it is possible to use a corpus which was not specifically designed to answer the current research questions.

The frequency of occurrence of plural /s/ is complemented by the observation that plural /s/ is not entirely constrained by its grammatical context. As discussed in Section 3.2.2.2, examining a morpheme which is less constrained by context means that there is the potential for higher uncertainty given the context, and thus a greater potential for the plural /s/ to be used to reduce uncertainty.

The final two reasons for selecting plural /s/ are first, that it is a single-segment morpheme, and second, that it has been shown to have the longest average duration and largest standard deviation in duration of any English /s/ morpheme (American English: Plag et al., 2017; NZE: Zimmermann, 2016, personal communication). Since it is a single-segment morpheme, if there are effects of predictability on the phonetic realization of the morpheme, these effects must be manifested in the /s/. Finally, a segment which on average has longer duration has more room for variation. A short segment could potentially get very long, but is bounded on the other end because there is a limit to how short a segment can be. A long segment, however, can get either shorter or longer depending on the need for signal specificity. Additionally, as plural /s/ has the greatest standard deviation in duration of any morphemic /s/, there is a large space of possible durations in which to find subtle differences related to predictability.

3.3.2 Corpus

The data for this study comes from the Origins of New Zealand English (ONZE) corpora (Gordon, Maclagan, & Hay, 2007), housed at the New Zealand Institute of Language, Brain and Behaviour (NZILBB) at the University of Canterbury. In total, the corpora used for this study contain over 1000 hours of recorded speech. All of the corpora in the ONZE collection consist of recordings of spoken English, collected at different times throughout the history of New Zealand. The first corpus, the Mobile Unit (MU) corpus, consists of interviews with New Zealanders born between 1851 and 1910, and was collected by the NZ National Broadcasting Service in 1946-8. The second corpus, the Intermediate Archive (IA), consists of interviews with speakers of New Zealand English born between 1890 and 1930, recorded between 1960 and the 1990s. Finally, the third corpus, the Canterbury Corpus (CC), consists of interviews and lists of words collected by Linguistics students at the University of Canterbury beginning in

1994. It is balanced for gender, age, and social class, and includes speakers born between 1930 and 1984.

3.3.3 Key Factor: morphological predictability

In examining the duration of plural /s/, many control factors are included (see Section 3.3.4), which allow the analysis to account for known sources of variation and also monitor whether the dataset is behaving as expected with regard to factors that have been previously shown to affect segment duration. By accounting for known sources of variation, the model is better able to test whether the key factor, syntagmatic morphological predictability, influences plural /s/ duration.

Syntagmatic morphological predictability is operationalized here using a measure of contextual predictability based on the preceding word: *log*[25] *preceding word plural probability* (PWPP, range: -8.053 to -1.063; mean: -3.21; sd: 1.24). This factor was measured by calculating, for each word preceding one of the plural words, how often this preceding word occurs before a plural relative to its total frequency in the corpora. Note that this corresponds to the Shannon Information (Shannon, 1948) for plurality, when multiplied by -1. In the equation below, PL refers to any plural word:[26]

$$PWPP(w_i) = \log \frac{freq(w_{i-1} + PL)}{freq(w_{i-1})}$$

For example, the word *various* occurs frequently before plural nouns, and has a PWPP of -1.063, while *pretty* has a much lower PWPP of -6.333. This allows us to quantify the contextual probability of the plural morpheme independently of the contextual probability of the word as a whole. PWPP ranges from -8.053 to -1.063, which means that the preceding word least likely to be followed by a plural has a probability of 0.000318 before taking the log, while the word most likely to be followed by a plural has a probability of 0.345. Note that this maximum does not approach a probability of 1,

[25] All logged measures in this paper use the natural log, the default in R.

[26] This measure is calculated based on our estimated list of plural words in the ONZE corpora. Because the corpora are not syntactically tagged, this is an imperfect measure but is consistent across the dataset. The list of plural words across which this is calculated includes all *s*-final words which are most frequently plural nouns according to CELEX (Baayen, Piepenbrock, & Gulikers, 1995), as well as irregular plural nouns.

as the probability was calculated across all instances of the preceding word, regardless of whether or not it was followed by a noun.

PWPP was calculated over all occurrences of the preceding word and also over only instances where the preceding word occurred before a noun.[27] When calculated over just nouns, the range of probabilities is from 0.00188 to 0.667. However, these two measures behaved similarly in the analysis and are highly correlated ($r_s = 0.96$, $p < .01$), so only effects of overall PWPP are reported here.

Additional steps were taken to ensure that any effect of PWPP was not driven by a small subset of previous words. The 10 previous words which occur most frequently in the data set (*the, of, and, three, two, these, six, few, other, their*) account for 41.3% of the total tokens, but have a wide range of PWPP scores (-4.74 to -1.27). The models presented below were run with and without these 10 most frequent previous words, and the results remained essentially the same, indicating that these most frequent words were not driving the effect.

A measure of following word plural predictability (the predictability of plurality, given one following word) was also calculated, but was not found to be predictive. This is somewhat surprising, given that much of the research related to word bigram predictability has found predictability based on the following word to be more important than the preceding word (see Bell et al., 2009). This will be discussed further in Section 3.5.

3.3.3.1 Paradigmatic morphological predictability

While the focus of this study is syntagmatic morphological predictability, a measure of paradigmatic predictability, the relative frequency (RF) of the complex word to the base word, is also included as a predictor (range: 0.51 – 7.23; mean: 0.94; sd: 0.31). Cohen (2014) finds that a higher log RF is correlated with longer 3sg /s/ duration in English, but shorter duration of the base. For derivational affixes, there have been conflicting previous findings. Seyfarth (2016) finds no effect of RF on 3sg /s/ duration in English. Hay (2004) finds that deletion of [t] is more likely in words like *swiftly* when the complex word is more frequent than the base. Schuppler et al. (2012), on the other hand,

[27] Again, because the corpora are not syntactically tagged, this measure was not exact. Any word which could be a noun, according to CELEX, was included.

find that [t]-deletion was less likely in Dutch verbs when the complex word is more frequent than the base. Cohen suggests that this may be due to different effects on the affix and the base.

MOP also makes different predictions about the affix and the base. For effects on the affix, making a prediction based on the MOP framework would require treating the target plural word as the context. In this case, if the RF of the complex word to the base word is higher, the message predictability (based on treating the target word as the context) is higher, and so the plural /s/ will be shorter.[28] This would allow for conserving resource cost in a case where the higher message predictability means higher probability of successful message transmission. However, this is not what Cohen found.

On the contrary, the RF of complex word to base word does not provide information about the predictability of the base, which would be required in order to make predictions about reduction of segments in the base. Therefore, according to MOP, if the base and affix are treated as separate units, the likelihood of deletion or weakening in the base is not influenced by RF. Instead, a measure of the predictability of the base given the affix would be required. For example, given a plural /s/, how likely is the base word *cat* as compared to *parrot*.

However, the effects seen on bases in previous work might be accounted for by MOP, if different processing models were taken into account. According to Hay (2004), complex words with higher RF are more likely to be processed as whole words, rather than as a combination of base + affix. This might mean that, for words with low RF, identifying the base as an independent message-bearing unit is more important than for words with higher RF. However, the implications of different processing models on the predictions of MOP related to weakening of base segments is beyond the scope of this chapter, as it is focussed on the affix.

[28] Relative frequency is typically used when the expected effect is part of the representation of the complex word (see Hay, 2004), rather than an online effect. It is also typically used with derivational suffixes, which tend to differ more in their degree of decomposability (see Chapter 1) than inflectional affixes do. For this reason, RF may not influence the duration of plural /s/.

3.3.4 Control Factors

Control factors are split into three subsets: those related to predictability, extra-linguistic factors, and phonological factors.

3.3.4.1 Predictability-related factors

The overall predictability of a linguistic unit is some combination of context-free predictability (e.g. frequency) and context-sensitive predictability (e.g. conditional probability). Both types of predictability are controlled for, as detailed in this section.

3.3.4.1.1 Word frequency

Although word frequency is only one measure of predictability, which does not take context into account, it has long been shown to have effects on the realizations of words, segments, and even morphemes. In general, units in more frequent words display greater degrees of reduction than those in less frequent words (Bybee, 2001; Jurafsky et al., 2001; Pluymaekers et al., 2005b). However, not all studies which consider word frequency find it to be a significant predictor of variation in duration (Plag et al., 2017; Seyfarth, 2014). For example, Seyfarth (2014) looks at conversational speech and finds significant effects of previous- and following-word-bigram probability, as well as informativity based on the following word, but finds no effect of frequency when those factors are included.[29]

In this paper, word frequency is operationalized using log transformed word probabilities based on a modified Kneser-Ney smoothed language model, calculated from the entire (2.1 million word) ONZE corpus (MU, IA, CC). Both lemma (range: 1.10 – 6.18, mean: 5.12; sd: 0.98) and wordform (range: 3.05 – 5.33; mean: 4.69; sd: 0.56) frequencies were tested. If there are frequency effects, more frequent words are expected to have shorter /s/ duration.[30] This is because more frequent words are more

[29] A recent manuscript by Cohen Priva (in prep) suggests that spurious effects of frequency may emerge in datasets when in fact the underlying effect is due to contextual predictability or informativity. However, the reverse is not true. This suggests that in previous studies which find effects of frequency, but do not control for contextual predictability or informativity, the effect may have spuriously emerged as a result of one of these other factors.

[30] This prediction only holds for wordform frequencies, or for lemma frequencies under a decompositional account of word access. For a whole-word approach to word access, effects would not be expected for lemma frequencies.

predictable, thus the overall message predictability of the word is higher, and the signal specificity can be lower. However, because base duration is included in the model, in order for word frequency to have an effect, it would have to influence the plural /s/ above and beyond any effects on the whole word.

3.3.4.1.2 Contextual word predictability

The predictability of a word, given the words that surround it, has been shown to have an effect on reduction, both of the word as a whole and of segments within the word (Gahl, Yao, & Johnson, 2012; Jurafsky et al., 2001; Pluymaekers et al., 2005a; Seyfarth, 2014). Nevertheless, some studies do not find significant effects of contextual predictability on reduction (Jurafsky et al., 2001). Raymond et al. (2006) find that predictability based on the following word is predictive of t/d-deletion in content words but not function words, while predictability based on the preceding word is not predictive of deletion.

While results vary somewhat according to the phenomenon, directionality, type of word, and measure of predictability, all positive results in this area point to increased reduction in words which are more predictable, giving the surrounding words. If there are effects of either preceding or following word contextual predictability in this study, /s/ in words which are more predictable in context is expected to be shorter.

This study operationalizes contextual word predictability using the conditional probability of a word based on either the preceding or following word. These measures are calculated by taking the frequency with which the preceding/following word occurs with the target word, then dividing by the overall frequency of the preceding/following word. These measures are referred to as *preceding-word-bigram predictability* (range: 1.50 – 5.20; mean: 2.74; sd: 1.02) and *following-word-bigram predictability* (range: 2.29 – 5.34; mean: 3.39; sd: 0.94). The formula below shows the calculation of bigram predictability based on the preceding word.

$$P(w_i|w_{i-1}) = \frac{freq(w_{i-1} + w_i)}{freq(w_{i-1})}$$

Preceding-word-bigram predictability, *P(w|w-1),* is important not only because it has been shown to affect the realization of words and segments, but also because it is a potential confound for the target measure of morphological predictability, measured in terms of preceding word plural probability (PWPP, see Section 3.3.3). The difference

between $P(w_i|w_{i-1})$ and PWPP is that $P(w_i|w_{i-1})$ quantifies the probability of the *particular* target word given the preceding word, while PWPP quantifies the probability of any *plural* word given the preceding word. If there are effects of PWPP even when $P(w_i|w_{i-1})$ is included, there can be higher confidence that the effects are genuinely attributable to PWPP.

3.3.4.1.3 Average word predictability

In addition to local measures of bigram predictability, there is evidence that the average predictability (informativity) of a word affects its duration. Cohen Priva (2008, 2012, 2015) uses *segment* informativity, or the average predictability of a segment given the preceding and following segments, and finds that segments with higher informativity tend to be longer and are less likely to be deleted. This measure is not tested here. In the domain of *word* informativity, i.e. a word's average contextual predictability, Seyfarth (2014) shows that words which tend to occur in highly probable contexts have overall shorter durations, even when they occur in less probable contexts. Likewise, Piantadosi et al. (2011) find that cross-linguistically, word lengths (measured in orthography, syllables, or number of phonemes) are shorter for words with lower informativity. These results suggest that the representations of words are sensitive to the contexts in which they occur. While Seyfarth's effect was on whole word duration and not single segment duration, the two factors of preceding-word-bigram informativity (range: 1.50 – 5.20; mean: 2.74; sd: 0.66) and following-word-bigram informativity (range: 2.29 – 5.34; mean: 3.39; sd: 0.69) are considered. If there is an effect of word informativity, the expectation is that /s/ duration will be longer in words which have higher informativity.

3.3.4.1.4 Average PWPP

Given the findings related to the average word or segment predictability (Cohen Priva, 2008, 2012, 2015; Seyfarth, 2014), this study also investigates whether there is an effect on plural /s/ duration of the average PWPP of a given plural word. This is referred to as PWPP informativity, and is calculated by taking the average PWPP across all instances of a given plural word (range: -6.80 – -1.27; mean: -3.21; sd: 0.57). If there is an influence of PWPP informativity, the expectation is that higher PWPP informativity correlates with longer /s/ duration.

3.3.4.2 Extra-linguistic factors

3.3.4.2.1 Speech rate

Speech rate is calculated as the number of syllables per second over the utterance containing the target word, excluding the target word itself. Utterance is defined as in the ONZE corpus (mean utterance length = 21 words). The domain of the 'utterance' is somewhat loosely defined in ONZE: the transcription guide instructs transcribers to "start each major utterance with a breakpoint." Speech rate is log transformed and used as a predictor (range: 0.83 – 2.57; mean: 1.70; sd: 0.22). Faster speech rates are predicted to correlate with shorter /s/ duration. This follows from studies which show increased deletion of segments at faster speech rates (Fosler-Lussier & Morgan, 1999; Raymond et al., 2006), shorter segment duration in faster speech (Byrd & Tan, 1996), and shorter word and suffix duration in faster speech (Pluymaekers et al., 2005a). Note that, as the duration of the base is included as a factor, a significant effect of speech rate on /s/ duration would mean that speech rate affects the plural /s/ above and beyond effects on the rest of the word.

3.3.4.2.2 Base duration

Because one of the goals of this study is to determine whether morphemes vary independently of the words to which they are attached, the log-transformed duration of the target word without the plural morpheme (*base duration*; measured in seconds) is also included as a factor (range: -2.04 – -0.69; mean: -2.31; sd: 0.29). If the duration of the plural morpheme is affected by the key factor even when controlling for base duration, it will demonstrate that the morpheme is subject to reduction or enhancement independently of the whole word.

3.3.4.3 Phonological factors

3.3.4.3.1 Plural type

The plural is coded based on whether it is phonologically [s] or [z] ([ɪz] plurals are excluded from this dataset). Voiced fricatives tend to be shorter than voiceless fricatives in English. This is supported by Seyfarth, Buz, and Jaeger (2016), who find that phonological [z] tends to be shorter than phonological [s], as well as Plag et al. (2017), who find that across all types of word-final /s/ in English, voiced tokens were shorter than voiceless tokens (coded acoustically). Consequently, all else being equal, [z] tokens are expected to have shorter duration than [s] tokens.

3.3.4.3.2 Phonological environment

The surrounding phonological environment is well known to influence segment variability (Fasold, 1972; Guy, 1980; Guy, 1991; Guy & Boberg, 1997; Labov, 1968; Labov, 1972; Pluymaekers et al., 2005a; Raymond et al., 2006; Zue & Laferrière, 1979). For example, Raymond et al. (2006) find that the likelihood of t/d deletion in English is higher when the preceding consonant is a homorganic sonorant, but not a homorganic obstruent. On the contrary, Guy (1980) finds that preceding sonorants disfavor t/d deletion, but does not take syllable position into account. Regarding /s/ specifically, Plag et al. (2017) find evidence that a following pause correlates with longer duration, but there was no significant effect of the manner of following segments when there was no pause. Plag et al. include preceding manner in terms of vowel vs. consonant, but do not distinguish different manners of articulation of consonants as a factor. They find that /s/ after a vowel is longer than /s/ after a consonant.

In this study the place and manner of articulation of the phone preceding the plural /s/ and the phone following the plural /s/ are included as control factors. The environment is controlled to some extent by including only consonant-final base words. Plural words for which the following phone is a strident were also excluded due to measurement difficulties related to differentiating between the target /s/ and the following strident when automatically extracting duration.

The factors of following place and manner of articulation also include 'pause' as one level of the factor, in order to account for effects of utterance-finality. Lengthening of words and segments which are phrase- or utterance-final has been shown in many studies (e.g. Beckman & Edwards, 1990; Bell et al., 2003; Cambier-Langeveld, 2000; Campbell & Isard, 1991; Fougeron & Keating, 1997; Klatt, 1976). In the present dataset, words are coded for whether or not they are at the end of an utterance, as transcribed in ONZE. Utterance-initial words were excluded, as the factor of interest relates to the preceding word. Utterance-final words are expected to have greater /s/ duration.

3.3.4.3.3 Similarity avoidance and the OCP

The Obligatory Contour Principle (OCP), originally proposed by Leben (1973), was formulated as a categorical constraint prohibiting identical underlying sequences of tones. It was extended to segmental phonology, accounting for restrictions against the occurrence of adjacent identical segments or features (McCarthy, 1986; Yip, 1988).

The OCP has traditionally been treated as a categorical constraint. There is reason to challenge this assumption, however. Odden (1986) shows that a strong version of the constraint is subject to many counterexamples, and scholars including Yip (1988) and Borowski (1987) raise questions about the degree of similarity required in order for segments to be subject to the OCP. Building on these challenges, other work considers treating the OCP as a tendency to avoid similar segments rather than a categorical constraint.

Long-distance effects of the OCP have been documented for many languages (e.g. Graff & Jaeger, 2009; McCarthy, 1986; Odden, 1986; Yip, 1988), including English (Berkley, 1994; Davis, 1991; Hay, Pierrehumbert, & Beckman, 2004; Pierrehumbert, 1993, 1994). Frisch, Pierrehumbert, and Broe (2004) show for Arabic that the conditioning factors can be gradient; that is, violations occur more or less often depending on the specific phonological violation. While identical consonants almost never co-occur in Arabic roots, the frequency of co-occurrence of non-identical segments is also restricted, as a function of their degree of similarity. However, Graff and Jaeger (2009) examine Aymaric, Dutch, and Javanese, and find that a measure of similarity that is feature-specific and sensitive to the distance between consonants is a better predictor of co-occurrence patterns than the measure used in Frisch et al.

Finally, coronal obstruents have been shown to be subject to OCP effects in English (Berkley, 1994; Hay et al., 2004; Pierrehumbert, 1994; Yip, 1988). In fact, Yip (1988) makes specific claims about the OCP with regard to plural /s/ in English. Yip suggests that for level 1 suffixes (e.g. *–tion*, *-tive*), a categorical OCP constraint applies to any sequence of coronals, such that any base ending in a coronal will be suffixed with either the *–ion/–ive* or *–ition/–itive* allomorph of these suffixes (for example: *deduce* ~ *deduction* vs. *complete* ~ *completion*, *define* ~ *definition*). However, for level 2 affixes (e.g. *–s*, *–ed*), a categorical OCP constraint only applies when the two coronal consonants agree in terms of the features strident and continuant. This is used to explain plural forms such as *horses*, which takes the [ɪz] allomorph, while *bats* is permitted. It is unclear whether this pattern will extend to long-distance gradient OCP effects involving coronal consonants.

Thus far, evidence for OCP effects has come from phonological patterns.[31] However, given the phonetic origins of such patterns, one might expect to observe OCP effects manifested through phonetically gradient realizations of phonemes. This study examines the effects of onset and coda consonants with varying degrees of similarity to /s/ on the duration of word-final plural /s/, using both binary (presence/absence of a similar segment) and gradient (as in Frisch et al., 2004) metrics to assess the similarity of /s/ to segments in the word onset and coda.

The binary measures for word onset include the presence or absence of any coronal consonant in the onset, the presence or absence of a coronal obstruent in the onset, and the presence or absence of a coronal strident in the onset. For the coda, only the presence or absence of either a coronal obstruent or any coronal were tested, because there are few stridents in codas in this dataset. Gradient similarity scores between segments were calculated following Frisch et al. (2004),[32] taking into account the degree of similarity between two consonants. Scores were calculated using an online similarity calculator (Albright, 2006), with the feature matrix from Frisch (1997). Because the plural morpheme can be realized as [s] or [z], two scores were calculated for each segment in the onset or coda: one comparing each segment in the onset/coda to [s] and one to [z]. For each word, the final similarity score was calculated by taking the scores comparing each consonant to the actual allomorph of the plural in that word ([s] or [z]), and then, for words with complex onsets or codas, choosing the highest similarity score. These log-transformed scores will be referred to as the Onset Similarity Score (OSS; range: -3.53 – 0; mean: -2.03; sd: 0.82) and Coda Similarity Score (CSS; range:-2.03 – 0; mean: -1.39; sd: 0.48).

The prediction is that for one or more of these measures of onset or coda similarity, /s/ will be shorter when a similar consonant is present in the onset (long-distance effect) or coda.

---

[31] There is, however, related work on similarity effects in phonological encoding, with slower speech rates observed when there is phonological overlap in the onset of two consecutive words (see Jaeger et al., 2012 for review).

[32] Graff and Jaeger (2009) use a different measure of gradient consonant similarity, and suggest that the method used by Frisch et al. is not easily extendable to other languages. Further study of OCP effects including bound morphemes should consider other methods for measuring similarity.

3.3.4.3.4 Duration maintenance effects

Previous research demonstrates that there is a tendency for syllable rhymes to have similar durations regardless of how many segments they have. This finding also applies to entire monosyllabic words (Lehiste, 1970). This suggests that in words with fewer segments, individual segment duration should be longer. In fact, Klatt (1976) finds that in onsets, consonants tend to have shorter durations when they occur in clusters, and Plag et al. (2017) find that English /s/ duration (both absolute, and relative to base duration) decreased as the number of consonants in the coda increased, across all types of /s/.

Given this previous work, plural /s/ duration in words with complex onsets or complex codas is expected to be shorter than in words with simple onsets or simple codas, due to the greater number of segments in the word.

Lehiste (1970) also finds that, across productions of a given word, if one segment is produced with longer than average duration, other segments are produced with shorter than average duration in order to compensate. If this line of research regarding relatively constant duration is extended to apply across words, words with longer vowels should also have shorter /s/ duration, in order to maintain relatively constant syllable duration across words. In the present study, base vowels are coded as long ([aː ɔː ɜː iː uː eɪ aɪ ɔɪ əʊ aʊ]) or short ([æ ɛ ɪ ɒ ʊ ʌ]). Words with long vowels are expected to have shorter /s/ duration.

3.3.5 Data

As stated above, the data for this study is drawn from the ONZE corpora (Gordon et al., 2007). The current dataset was compiled by extracting all s-final words from three of the ONZE collections (MU, IA, CC), along with automatically extracted duration measures for each word and each /s/. These collections are discussed in more detail in Section 3.3.2. The dataset was then automatically filtered to exclude words which cannot be nouns (according to CELEX, Baayen et al., 1995).

The ONZE corpora are not syntactically tagged due to the difficulty of using automatic taggers on naturally occurring speech. Because of this, it was difficult to ensure that all and only plural nouns were included in the dataset. A decision was made to err on the conservative side, excluding any s-final words that have a reasonable chance of being non-nouns. To do this, only words for which the most frequent category is noun,

according to CELEX (Baayen et al., 1995) were kept. Additionally, any words for which the lemmatized form and the s-final form are the same were excluded, as another way of eliminating non-plurals (e.g. *trousers, linguistics*).

For the present study, the dataset was further limited to include only monosyllabic words ending with an [s] or [z] plural allomorph, where the [s]/[z] was preceded by a consonant and not followed by a sibilant, in order to control the phonological environment as much as possible. Note that due to NZE being a generally non-rhotic dialect, as well as frequent vocalization of post-vocalic /l/ in NZE (Hay, Maclagan, & Gordon, 2008), base nouns ending in post-vocalic /ɹ/ or /l/ were treated as potentially vowel-final, and thus were not included in the analysis. In addition to excluding non-nouns, words with obvious measurement errors were excluded, such as cases where the automatically measured plural duration was longer than the word duration, or the speech rate was extremely high. Values for speech rate which were more than 2.5 standard deviations above the mean were excluded. Additionally, because duration measurements were extracted automatically, outliers of plural /s/ duration were excluded. At the high end, plural durations which were more than 2.5 standard deviations above the mean were excluded. At the low end, plural durations at or below .03 seconds were excluded. Examination of the distribution of plural lengths showed a clearly bimodal distribution, indicating analysis errors. In particular, many tokens that were automatically assigned a duration of .03 seconds actually had no discernible /s/, which indicates that .03 may be some sort of default value that is assigned when the actual value is smaller than this, or difficult to detect. For this reason, the large cluster of points at exactly .03 seconds, and those below this threshold, were excluded. Finally, tokens for which either the target word or the preceding word had a frequency of less than 20 in the ONZE corpora were excluded due to the decreased reliability of frequency counts for low-frequency words. This restriction was applied to both target word and preceding word frequencies because preceding word frequency was used in the calculation of PWPP, the key measure of morphological predictability (see Section 3.2.1). PWPP scores are highly unstable at low previous word frequency counts, especially for words with low PWPP. While the cut-off of 20 was chosen somewhat arbitrarily, higher cut-offs were also tested (50, 100), and all effects remained qualitatively the same. This dataset was then hand-checked to ensure that each token included was indeed a plural noun, and those that were not were excluded. After these exclusions, 5275 tokens were analyzed (292 plural word types, 491 speakers).

The exclusions listed above are summarized in Table 3.1, along with the number of tokens which were excluded at each stage, and the percentages of the total 1-syllable words that they represent.

Table 3.1: Number of Tokens Excluded at Each Step, and Percentage of Total. Bolded Lines Show Total Exclusions for Each Category.

Exclusion Criterion	Tokens	% of Total
Non-target words	**876**	**5.08%**
Quantifiers (*heaps, lots, sorts*)	583	3.38%
Plural same as singular (e.g. *clothes*)	293	1.70%
Phonological context	**6866**	**39.81%**
Preceded by vowel	4681	27.14%
Preceded by postvocalic /l/	1483	8.60%
Followed by strident	346	2.01%
Utterance initial	356	2.06%
Measurement error / Missing data	**3013**	**17.47%**
Missing predictability data	1505	8.73%
Missing speech rate	273	1.58%
Word duration < plural duration	96	0.56%
Speech rate > 2.5 sds above mean:	8	0.05%
Plural length > 2.5 sds above mean:	193	1.12%
Plural length <= .03:	938	5.44%
Frequency cut-offs	**723**	**4.19%**
Target word frequency < 20	375	2.17%
Previous word frequency < 20	348	2.02%
Manual exclusions:	**496**	**2.88%**
Total tokens remaining:	**5275**	**30.58%**

3.3.6 Analysis

Effects of various factors on log /s/ duration (range: -3.22 – -1.41; mean: -2.41; sd: 0.44) were analyzed using linear mixed effects models (e.g. Baayen, 2008), using the lme4 package in R (Bates, Maechler, Bolker, & Walker, 2015; R Core Team, 2015). Models were fit using backwards elimination, starting with all possible factors included in the model and then, after testing for theoretically motivated interactions, eliminating one by one non-significant factors which contributed the least to explaining variance.[33] The total number of parameters tested is approximately 118 (26 for fixed effects, 6 for

[33] While this approach is commonly used to analyze corpus data, it is known to be anti-conservative, potentially resulting in inflated p values. However, it is preferred over forward stepwise selection (Harrell, 2001).

random effects, 1 error term, and 85 from interactions of fixed effects). This is well below the recommended maximum number of parameters allowable based on the number of observations (5275 / 15 = 351; see Jaeger 2011). Non-significant factors were initially identified by t values less that 2. ANOVA tests between models and comparisons of Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) scores were used to ensure that removing a factor did not significantly ($p < .05$) decrease the explanatory power of the model. Interactions were kept if they did significantly improve the fit of the model, as determined by examining AIC and BIC scores and conducting ANOVA tests. Preceding-word-bigram predictability was retained throughout, though it does not contribute significantly to explaining variation, because of the potential confound with PWPP. Random intercepts for speaker and base word[34] were included, as well as random slopes for PWPP by speaker and base word. The random intercepts account for individual differences in average /s/ duration for each speaker and each base word. The random slopes account for potential differences in the degree to which PWPP influences duration for different speakers or base words.

Because of strong correlations between all measures of onset similarity, the process was repeated separately for each of these measures, with the final selected model using the measure which contributed the most to explaining variance. While the gradient measure, Onset Similarity Score (OSS), was highly significant, further testing indicated that this effect was driven by the difference between the presence or absence of a coronal obstruent. When subsets of the data including or excluding coronal obstruents were tested independently, OSS was not a significant predictor.[35] Thus, the measure of onset similarity included in the final model is the presence or absence of a coronal obstruent in the onset.

The same process was carried out for the three measures of coda similarity and two measures of word frequency (lemma and wordform frequency). For coda similarity, the

[34] Note that the random intercept for base word is correlated with the following factors: plural type, place and manner of articulation of the preceding consonant, onset and coda complexity, vowel length, and word frequency. However, these other factors are included because they may have effects which are not captured entirely by base random intercepts, especially for low-frequency bases.

[35] While this is taken to indicate that the effect of OSS is driven by the presence of coronal obstruents, the absence of an effect in the data subsets could be due to having less power in these smaller datasets.

presence of any coronal consonant in the coda was significant, but the gradient measure was more a predictive factor. In order to verify that this effect was not entirely carried by the presence or absence of coronal consonants, the dataset was split into those words containing coronal consonants in the coda and those without. The effect remained significant in each dataset, indicating that there is indeed a gradient effect.

However, for word frequency, neither of the measures significantly contributed to explaining variation.

In order to test whether there is multicollinearity, the diagnostic Variance Inflation Factor (VIF; code from Franks, 2011) was used in order to test whether any two factors were highly correlated. A VIF score was calculated for each factor used in the final model, and all VIF scores were below 3. According to Zuur et al. (2010), this is an acceptable threshold.

3.4 Results

As detailed above, the optimal model was selected using backwards elimination. In the final model (shown in Table 3.2), the key prediction was confirmed, and many control factors showed the expected effects.

3.4.1 Key factor – morphological predictability

Regarding the key factor of morphological predictability, there was a significant main effect of preceding word plural predictability (PWPP), which can be seen in Figure 3.2. Higher PWPP correlates with shorter /s/ duration; in other words, the more probable the plural is given the preceding word, the shorter the duration of the /s/.

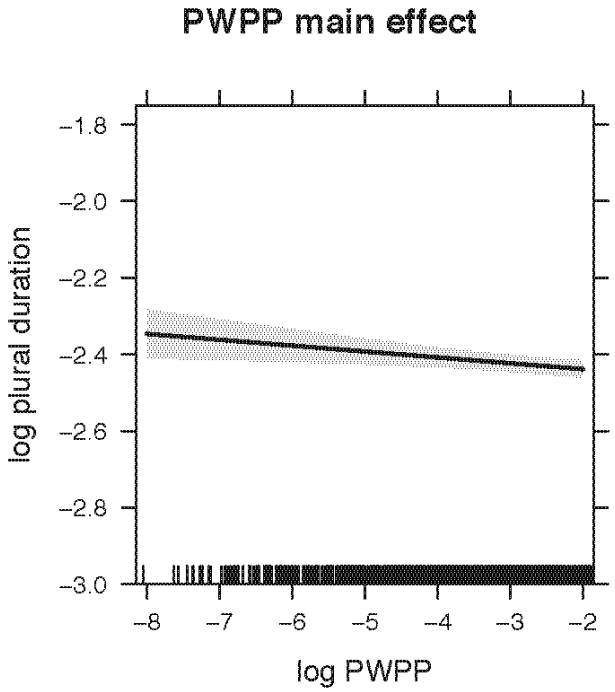

Figure 3.2: Key factor main effect.

3.4.2 Control factors

Numerous control factors also emerged as significant, both in interactions and as main effects. There is a significant interaction of speech rate with the gradient coda similarity score (CSS), as well as significant main effects of a coronal consonant in the onset, plural type ([s]/[z]), base coda and onset complexity, vowel length, preceding and following phone manner, log base duration, and corpus. Results are presented in three subsections, according to the domain of the control factors.

3.4.2.1 Predictability-related factors

Apart from morphological predictability, none of the factors related to predictability emerged as significant predictors of plural /s/ duration. Factors which did not show significant effects include wordform and lemma frequency, word bigram predictability based on the preceding or following word, informativity based on the preceding or following word, relative frequency of plural to singular, and PWPP informativity.

Table 3.2: Model Summary (Fixed Effects)

	Estimate	Std. Error	t value
(Intercept)	-1.566	0.097	-16.210
PWPP	-0.016	0.006	-2.789
onset coronal obstruent: TRUE	-0.074	0.019	-3.855
coda gradient similarity score	0.679	0.239	2.839
coda gradient similarity score * log speech rate	-0.539	0.141	-3.818
log speech rate	-0.102	0.049	-2.095
log base duration	0.383	0.024	15.819
plural type: z	-0.126	0.021	-5.985
preceding phone manner: fricative	0.158	0.042	3.747
preceding phone manner: nasal	0.051	0.025	2.023
base onset: simple	0.053	0.021	2.553
base coda: simple	0.081	0.029	2.817
corpus: MU	-0.089	0.024	-3.674
vowel length: short	0.064	0.018	3.590
next phone manner: affricate	-0.101	0.079	-1.267
next phone manner: fricative	-0.198	0.021	-9.244
next phone manner: liquid	-0.218	0.026	-8.522
next phone manner: nasal	-0.070	0.033	-2.143
next phone manner: glide	-0.218	0.022	-10.031
next phone manner: stop	-0.218	0.022	-9.958
next phone manner: vowel	-0.259	0.017	-15.357
preceding-word-bigram predictability	0.010	0.007	1.420

3.4.2.2 Extralinguistic factors

There is a significant interaction of speech rate with CSS, as shown in Figure 3.3. In the left side of the figure, showing the interaction, the top and bottom lines correspond to the minimum and maximum speech rates, with intermediate lines showing each quartile of the data. This plot shows that overall, higher speech rate correlates with shorter /s/ duration. The right side of the figure will be discussed below. Regarding the other extra-linguistic factors, longer base duration correlates with longer /s/ duration, and the effect of corpus shows that tokens from the MU corpus have significantly shorter /s/ duration compared to the CC and IA corpora.

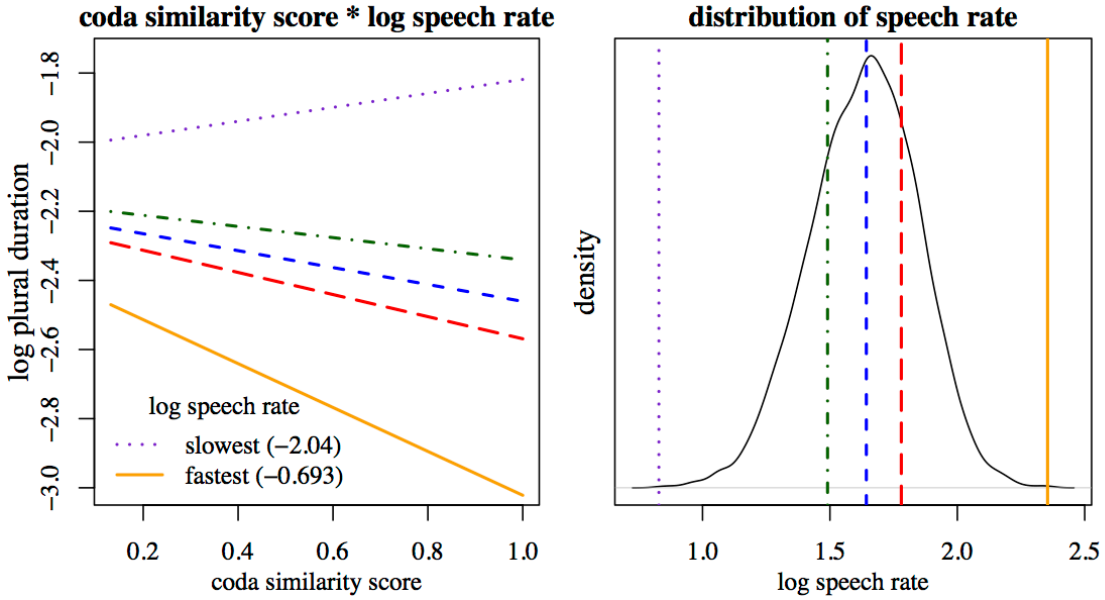

Figure 3.3: Interaction of speech rate and CSS.

3.4.2.3 Phonological factors

In terms of phonological factors, as discussed above, coda similarity score (CSS) interacts with speech rate (see Figure 3.3). As a reminder, CSS is a similarity score based on the degree of similarity between the coda consonants and /s/. As noted above, the left side of the figure shows the interaction of CSS and speech rate, and their combined influence on plural /s/ duration. The right side of the figure shows the distribution of speech rates across the dataset, demonstrating that the majority of the data occurs around the middle three lines. Thus, for the majority of the data, the effect of CSS goes in the expected direction, with /s/ being shorter when there is a more similar consonant in the coda.

There are also significant main effects of a coronal consonant in the onset, plural type ([s]/[z]), base coda and onset complexity, vowel length, and preceding and following phone manner. There is no effect of place of articulation of the surrounding segments. The effects of a coronal obstruent in the onset, plural type, syllable complexity, vowel length, and base duration are straightforward and go in the expected directions. When a coronal consonant is present in the base coda, /s/ duration tends to be shorter. Plural type [z] tends to have shorter duration, while simple onsets and codas correlate with longer /s/ duration. A short vowel in the base also correlates with longer /s/ duration.

Figure 3.4 shows the effects of preceding and following phone manner of articulation. The effect of following phone is driven by utterance-final lengthening, with a following pause correlating with the longest /s/ duration. Plural /s/ preceding a pause is

significantly longer than before all other segments, with the exception of affricates. Before a vowel, /s/ duration is significantly shorter than before any other segment except liquids. Before a nasal, /s/ is significantly longer than before fricatives, liquids, glides, stops, and vowels.

The effect of preceding phone manner only shows a significant difference between fricatives and the other manners, with no significant difference between stops and nasals. However, it is important to note that there are relatively few preceding fricatives, compared to other preceding phone types (n=376; compare to n=2894 for stops, n=2005 for nasals).

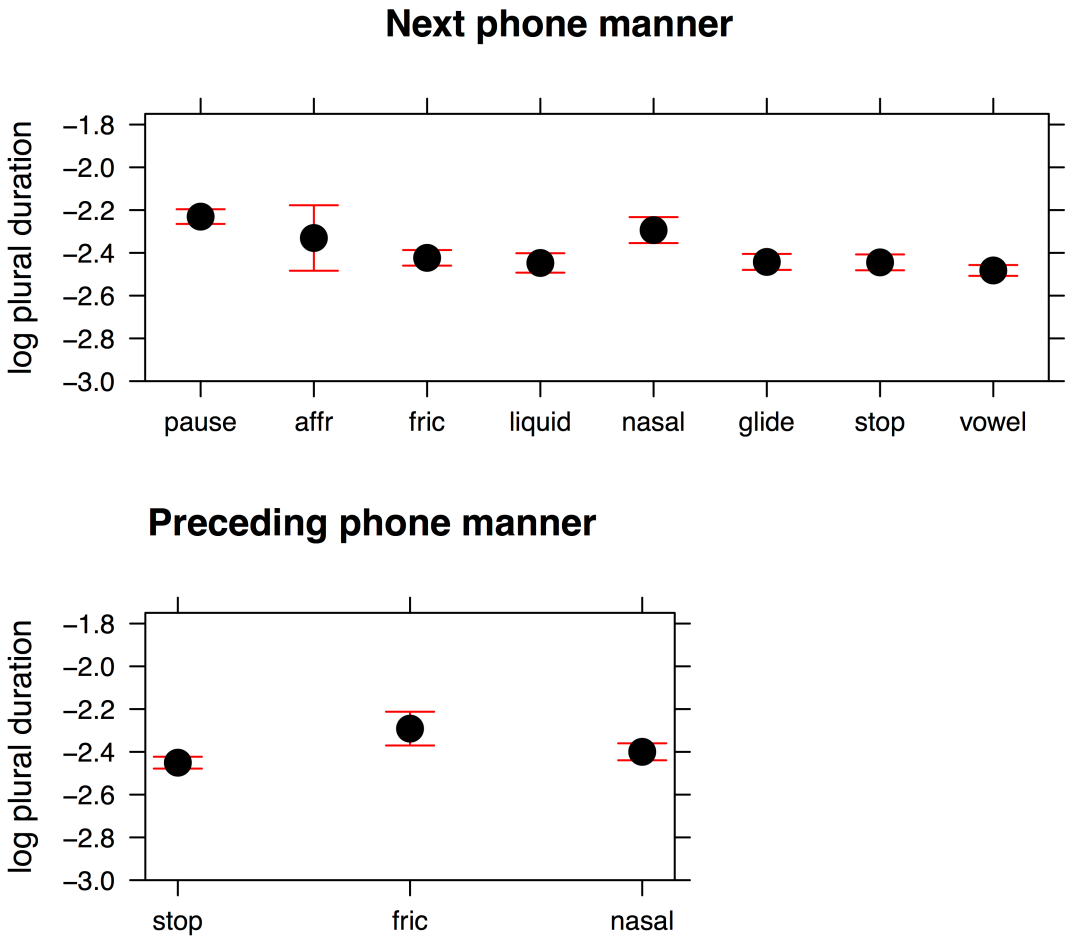

Figure 3.4: Surrounding phone manner effects

3.5 Discussion

3.5.1 Morphological predictability

The present study shows a significant effect of the hypothesized novel factor of morphological predictability on the phonetic realization of NZE plural /s/. The influence of the contextual predictability of plurality on plural duration confirms that predictability operates not only at the level of the word and the segment, but also at the level of the morpheme. This, in turn, suggests that morphological structure is relevant to the phonetic realization of morphemes. The duration of /s/ is modulated by its contribution to identifying *plurality* in context, not simply identifying the word. While /s/ is a single-segment morpheme (and therefore the only segment in the word explicitly signalling plurality[36]), its contribution to identifying plurality varies based on PWPP, the predictability of plurality given the context. In plural nouns preceded by a word which is highly predictive of plurality (e.g. *various,* PWPP= -1.062894), the results suggest that since /s/ is contributing less to identifying the noun as plural, it can be reduced. In this case, there is less likelihood that the message of plurality will not be successfully conveyed. In the terminology used in the MOP framework, when message predictability is higher, signal specificity and hence redundancy in the plural's signal can be reduced. On the other hand, when preceded by a word with low plural predictability, e.g. *pretty* (PWPP= -6.33328), more cues to the plural are needed and thus signal specificity is increased resulting in the increased duration of /s/. This result is consistent with findings showing that when words and segments are more predictable in context, they are more likely to be reduced (e.g., Bell et al., 2009; Raymond et al., 2006; Seyfarth, 2014). This result suggests that properties of sub-lexical meaningful units contribute to systematic phonetic variation. Note that, unlike the findings in Cohen (2014), this effect is consistent across all target words, regardless of wordform frequency. This finding supports Hypothesis 3:

H 3: Linguistic cues signaling plurality are produced with more reduced realizations when they are more predictable.

[36] There is some evidence that speakers can distinguish unaffixed and affixed forms of bases (Kemps et al., 2005a, 2005b), which means that there may be subtle phonetic cues to plurality earlier in the word.

The present study differs from previous studies in which effects of morphological structure are proposed to be due to a segment's contribution to identifying the word as a whole (e.g. Hanique & Ernestus, 2012). In the present case, systematic variation in the duration of plural /s/ is shown to occur even when accounting for across-word differences and word bigram predictability, indicating that the effect is at work within the same plural word in different contexts. Given that the relative contribution of a segment to identifying the whole word, independent of context, remains constant for a given word, this systematic variation must have another explanation, namely the variation in the predictability of the plural given the context. These findings thus contribute to the debate over whether or not morphological structure influences the phonetic realization of morphemes, suggesting that it does in fact play a role. This, in turn, weighs in on the debate over whether complex words are necessarily processed as entire units. If plural /s/ duration varies independently of base duration and is affected by the contextual predictability of plurality, this suggests that the plural morpheme can operate independently of the word in which it is contained. This finding provides support for Hypothesis 4:

H 4: Bound morphemes do have independent representations of some nature.

3.5.1.1 Morphological predictability based on the following word

The absence of an effect for the predictability of plurality given the following word is worth noting, particularly given previous results which suggest that, for word predictability, the following context is more important than the preceding context. Specifically, Bell et al. (2009) find a strong effect of following word conditional probability on word duration for both content and function words, but no effect of previous word conditional probability on content words. For function words, while preceding word bigram predictability is a significant predictor, it is only strong for very high frequency function words.

In the present study, previous word plural predictability is found to influence plural duration, while following word plural predictability is not. Given that plural words are content words, this suggests a difference in the nature of how predictability influences words compared to how it influences bound morphemes, which should be explored further. It appears that the nature of the predictability of bound morphemes is different than word predictability, in ways that still need to be determined. It could be that the stronger influence of preceding predictability is due to the fact that nouns are often at

the end of noun phrases, so the words which are most closely linked to them, and thus most predictive of plurality, tend to occur before the noun. It is also possible that a more constrained measure of following word morphological predictability might prove to be an important predictor. For example, if only plural nouns that are directly followed by a verb were examined, there may be a strong effect since verbs following nouns typically agree in terms of number specification.

3.5.2 Control factors

In addition to the effect of syntagmatic morphological predictability, the effects of the control factors plural type, coda and onset complexity, vowel length, base duration, and following phone manner all corroborate previous findings, as outlined in Section 3.3.4. However, many factors did not show significant effects, including wordform and lemma frequency, bigram predictability based on the preceding or following word, informativity based on the preceding or following word, relative frequency of plural to singular, and place of articulation of the surrounding segments.

3.5.2.1 Predictability-related factors

It is worth noting that none of the measures of predictability, apart from PWPP, emerged as significant. This is most likely due to the fact that base duration was included as a factor, so any effects of these predictors on whole word duration were accounted for. This suggests that factors such as wordform frequency and bigram word predictability do not have effects of /s/ duration independent of their effects on whole word duration. The absence of an effect of relative frequency of the complex word and the base is unexpected, given previous work showing such effects for bound morphemes (Cohen, 2014; Hay, 2004; Schuppler et al., 2012; see, however, Seyfarth, 2016). However, this may be due to differences between inflectional and derivational morphemes, or between experimental and corpus data.

3.5.2.2 Extralinguistic factors

In terms of extralinguistic factors, base duration has the predicted effect, with longer base duration correlated with longer /s/ duration. This indicates that if there is some factor influencing the duration of the base, this factor is also affecting the /s/, as a part of the whole word.

Overall, the effect of speech rate also goes in the expected direction, with higher speech rates correlating with shorter /s/ duration. This effect is stronger for words with higher

Coda Similarity Score (CSS), but is consistent across all words. The effect of speech rate on /s/ duration, even when controlling for base duration, shows that /s/ is reduced to a greater degree than the rest of the word. This could be related to the /s/ being word-final. On average, word-final segments tend to be less important in terms of identifying the word, thus are better targets for reduction (van Son & Pols, 2003; Hall et al., submitted).

3.5.2.3 Phonological

All of the phonological factors included as controls did emerge as significant predictors, with the exception of preceding and following phone place of articulation. With regard to phonological voicing, plurals which are phonologically [z] tend to be shorter than those which are phonologically [s], as expected, because voiced fricatives are in general shorter than voiceless fricatives (e.g. Plag et al., 2017; Seyfarth et al., 2016).

Regarding duration maintenance effects, the effects of coda complexity, vowel length, and onset complexity support the idea that syllable rhymes and whole words each tend to maintain similar durations across words, all else being equal (Klatt 1976; Lehiste 1970). A shorter vowel or fewer coda consonants make the initial part of the rhyme shorter, leaving room for the /s/ to be longer. Likewise, fewer consonants in the onset make onset duration shorter, leaving room for the /s/ to be longer, if overall word duration is constant, all else being equal.

In terms of phrase-final lengthening, the effects of following phone manner show that before a pause, /s/ tends to be longest. However, apart from a following pause, none of the other following manners correlate with significantly different /s/ duration.

3.5.2.3.1 Similarity avoidance and the OCP

The OCP effects related to both the onset and the coda are particularly interesting. A coronal obstruent in the onset predicts shorter /s/ duration, consistent with studies showing categorical effects of the OCP in English across intervening phonemes (Berkley, 1994; Hay et al., 2004; Pierrehumbert, 1994). However, the present results provide further evidence suggesting that the OCP operates even at the level of fine phonetic detail. In particular, in this dataset, the presence of a similar segment in the onset is correlated with gradient differences in duration, rather than the presence or absence of the similar segment, an effect that is commonly associated with the OCP (Berkley, 1994; Davis, 1991; Hay et al., 2004; Pierrehumbert, 1993, 1994). Although

the gradient measure of onset similarity (OSS, a measure of the degree of similarity between the onset consonants and /s/) was not shown to be more predictive than the binary factor of the presence of a coronal obstruent in the onset, the effect of the binary similarity factor on /s/ duration is a gradient phonetic effect of long-distance similarity avoidance, a phenomenon that has not previously been recorded in the literature.

The effect of the coda similarity score (CSS) gradient factor extends this novel OCP effect even further, showing that gradient degrees of similarity can have gradient effects on the phonetic realization of phonemes, at least when they are in close proximity. CSS was a better predictor of /s/ duration than a binary measure of the presence or absence of any coronal, and remained significant even when the data was split into tokens with coronal and non-coronal codas.

In terms of the domain and specificity of OCP effects, the similarity avoidance effects found regarding the onset and the coda show quite different patterns. The long-distance effect of similarity to consonants in the onset is sensitive to any coronal obstruent, rather than to any coronal consonant, or to strident consonants alone. The reduction of /s/ when co-occurring with a coronal obstruent seems to have a level of sensitivity in-between the categorical effects discussed by Yip (1988), regarding coronal suffixes in English. Yip finds that the distribution of allomorphs for level 1 coronal suffixes are sensitive to any coronal consonant, while the allomorphs of level 2 coronal suffixes (e.g. *–s, –ed*) are only sensitive to coronal consonants which agree in terms of the features [strident] and [continuant] (as discussed in Section 3.3.4.3.3).

The coda effect, however, extends across all phonemes, demonstrating a gradient similarity effect that is not limited to coronals. These differences in sensitivity between the onset and coda effects may simply be a result of the data having different numbers of coronals of various types in onset position, as opposed to coda position, or they may indicate some qualitative difference between the coda and onset. Future studies might further explore whether long-distance and local OCP effects are influenced by the same factors, but this is outside the scope of the present study.

On the whole, this study demonstrates that the production of the English plural is phonetically gradient, and varies independently of the base word. It is conditioned by both plural predictability in context, and several phonological and extralinguistic factors, including long-distance and local gradient effects of the Obligatory Contour Principle.

3.5.3 Conclusions

The study described in this chapter provides further evidence that morphemes can be treated as message-bearing units, and supports both Hypotheses 3 and 4. Plural morphemes which are more predictable, given the context, are produced with more reduced realizations, independently of word-level effects.

If these message-bearing units are influenced by message predictability, as calculated over one preceding word, this raises the question of whether the size of the relevant context is greater than one word. While the measure of message predictability used here is a reasonable approximation, just how much context is relevant when calculating the message predictability of a plural morpheme is an open question. Chapter 4 explores other ways of measuring the contextual predictability of plurals, both by obtaining plural predictability ratings directly from language users, and by using ratings from a larger preceding context.

4 How much context matters? Crowd sourcing ratings of plural contexts

4.1 Intro

In Chapter 3, a corpus study of plural /s/ duration in New Zealand English demonstrated that the duration of plural /s/ is influenced by the predictability of the message of plurality, given one preceding word (preceding word plural predictability; PWPP). This effect indicates that language users have knowledge of the statistical properties of bound morphemes, as hypothesized in Chapter 1. In the terminology of the Message-Oriented Phonology (MOP) framework, the signal specificity of the plural morpheme, and thus the resource cost, is lower when the message predictability is higher, and vice versa. This shows the predicted trade-off between resource cost and message transmission accuracy.

The finding that effects of message predictability based on one preceding word can be seen raises the question of just how much context is relevant to predicting the phonetic realization of bound morphemes (specifically New Zealand English plural /s/). In the MOP framework, as well as in other Bayesian approaches to linguistics, the context is essential. However, the type and size of the relevant context for each research question is an empirical question. For example, when calculating plural predictability, if a word

such as *are* occurs two or three words before a noun, this presumably makes a plural more probable because of the use of *are* as a copula linking two plural noun phrases (e.g. *we are old friends*). However, because the spoken corpus used in Chapter 3 is not syntactically tagged, it is difficult to reliably extract nearby verbs or other parts of speech which may be particularly helpful in predicting plurality. A measure of plural predictability which calculates probability based on a larger context might be able to capture situations such as this, where words earlier in the sentence influence predictability. Additionally, the study in Chapter 3 raises the question of whether there are other ways of estimating morphological predictability, apart from using frequency counts from a corpus.

This chapter extends the measure of morphological predictability beyond one preceding word, using a set of measures which will be called preceding *context* plural predictability (PCPP) scores. These PCPP scores are measures of morphological predictability based on judgments of preceding contexts of one or five words, collected in an online rating task from speakers of English. The motivation behind using this method is to more directly access the probabilities used by speakers, but also to be able to capture aspects of predictability that may not be captured well by *n*-gram probability models. By using contexts extracted from the same corpus used in Chapter 3, these ratings can then be tested to see whether they are predictive of plural duration. By collecting judgments based on preceding contexts of different sizes, several research questions can be addressed. While this study uses only two context sizes in order to test this methodology, future work could explore additional context sizes.

Although calculating a PCPP score over a larger context is possible from a corpus, there are two main reasons that subjective ratings were chosen. One is that there would be problems related to data sparsity if the ONZE corpora used in Chapter 3 were used to calculate PCPP scores over five words. Already in Chapter 3, where just one preceding word was used to calculate plural predictability, many preceding contexts had to be excluded because they rarely occurred before plural nouns. By increasing the number of preceding words included, these issues would be compounded. Alternatively, these predictabilities based on a larger context could be calculated from a larger corpus, such as the Web 1T 5-gram Corpus (Brants & Franz, 2006). However, this presents its own difficulties, including large computational costs, using a corpus which does not consist

of spoken data, and difficulties capturing the predictability of an abstract category such as plurality.

This difficulty in capturing the predictability of the abstract category of plurality with *n*-gram probability calculations is the second reason for using subjective ratings rather than corpus ratings. Even with a larger corpus, *n*-gram word probabilities capture the likelihood of a plural to follow an exact sequence of *n* words. While this is one way of estimating plural predictability, it assigns equal weighting to all words, and may result in artificially low predictability if the sequence includes even one rare word. For example, the sequences *Each one of the incredible* $NOUN_{PL}$ and *Each one of the unfathomable* $NOUN_{PL}$ would have very different n-gram plural predictabilities because the word *unfathomable* is much more rare than *incredible* (0.89 occurrences per million words vs. 19.39 according to CELEX, Baayen et al., 1995). On the contrary, a human rater would probably rate these two sequences fairly similarly in terms of plural predictability because it is much more likely that the sequence *Each one of the…* will be followed by a plural noun rather than a singular noun. This sensitivity to words or sequences of words which are particularly relevant to predicting plurality, and the corresponding insensitivity to words which are less relevant (e.g. *unfathomable*) is an important advantage to using subjective ratings.

Subjective ratings from language users have been used, with varying degrees of success, to approximate word frequency (e.g. Kuperman & Van Dyke, 2013) and to estimate whether certain words are used more by certain social groups such as older vs. younger speakers, or males vs. females (Kim, 2016; Walker & Hay, 2011, ms). Additionally, cloze probability ratings have been used to estimate the predictability of various linguistic structures or grammatical function assignments (e.g. Kurumada & Jaeger, 2015; Tily & Piantadosi, 2009). In terms of word frequency, subjective ratings have even been found to better predict behavior compared to frequency estimates from a corpus, which may overestimate speakers' familiarity with lower frequency items (Kuperman & Van Dyke, 2013). These studies will be discussed further in Section 4.2.1. The broad research questions addressed in this chapter are:

RQ 3: How is the production of linguistic cues which signal the grammatical category of plurality influenced by predictability?
RQ 5: Is this knowledge of statistical properties of morphological cues available at a conscious level?
RQ 6: What is the size of the context used to track the predictability of morphemes?

In this chapter, these questions will be addressed through the more specific research questions:

RQ 4.1: How do subjective ratings compare to corpus-based ratings of morphological predictability (for one-word preceding contexts)?
RQ 4.2: Does plural /s/ duration vary systematically based on subjective ratings of preceding contexts?
RQ 4.3: Do subjective ratings based on different amounts of context contribute in different ways to predicting /s/ duration?

Previous work comparing subjective and corpus-based ratings finds that these tend to be correlated (Kuperman & Van Dyke, 2013; Melnick, Jaeger, & Wasow, 2010; Walker & Hay, ms), suggesting that the PCPP ratings in the present study will be correlated with the corpus-based PWPP scores from Chapter 3 (RQ 4.1). If subjective ratings are another valid way of estimating morphological predictability, as they seem to be for word frequency (Kuperman & Van Dyke, 2013), the prediction is that plural /s/ duration will vary based on subjective ratings of preceding contexts, such that /s/ duration is shorter in more predictable contexts (RQ 4.2). These ratings might even be more predictive than corpus-based estimates, because they come directly from speakers (Kuperman & Van Dyke, 2013). This would indicate that language users do have conscious access to plural predictability. However, it is possible that these subjective ratings may not be predictive at all. Walker and Hay (2011) found that subjective ratings of which social group used a given word more did not interact significantly with speaker voice in a lexical decision task, while corpus-based measures of the association of social group with word did.

With regard to context size, if larger contexts are relevant to plural predictability, then ratings from larger contexts will more accurately capture morphological predictability, at least subjectively. If these larger contexts capture a more detailed representation of message predictability, then ratings based on these contexts should explain more variation in signal specificity (/s/ duration). Therefore, the prediction is that ratings based on larger contexts (five words) will be more predictive of plural duration than those based on smaller contexts (one word) (RQ 4.3). Future studies might explore even larger contexts, to test whether there is a limit on the size of the relevant context.

These predictions tie into the trade-off between message predictability and signal specificity discussed in Chapter 3. As the measure of contextual predictability becomes more nuanced, the expectation is that this trade-off will be able to be seen at even finer-

grained levels, unless it is the case that only the immediate context is relevant to the realization of bound morphemes. However, if subjective ratings are not in fact predictive of /s/ duration, this does not necessarily mean that the larger context is not relevant. Rather, it could suggest that this method of calculating morphological predictability does not capture the relevant information, and that language users do not have conscious access to morphological predictability.

The remainder of this chapter is organized as follows: Section 4.2 expands on the results of studies which use subjective ratings; Section 4.3 discusses the methods used for collection of ratings and analysis; Section 4.4 presents the results, while Section 4.5 discusses the findings and concludes the study by referring back to the overall theme of the thesis, the influence of morphological predictability on bound morphemes.

4.2 Background

4.2.1 Subjective vs. corpus ratings

There is a substantial body of work investigating the use of subjective ratings either alongside or in place of corpus-based frequencies to investigate variation across individual words (Kim, 2016; Kuperman & Van Dyke, 2013; Walker & Hay, 2011), as well as to predict usage of optional grammatical markers (Bresnan, 2007; Melnick et al., 2010). The majority of this work uses these subjective ratings as predictors for dependent variables in perception tasks, such as reaction time or error rate in lexical decision tasks, or eye movement in reading tasks. However, there is also research using cloze probability ratings to predict production patterns for both grammatical function assignment (e.g. Kurumada & Jaeger, 2015) and the realization of noun phrases (e.g. Kravtchenko, 2014; Tily & Piantadosi, 2009). For example, Tily and Piantadosi (2009) use cloze probability ratings of excerpts from the Wall Street Journal to estimate the likelihood of upcoming nominal referents. They find that in contexts where participants in the cloze probability task are more likely to correctly guess the referent (meaning the referent is more predictable), writers of the Wall Street Journal are more likely to use pronouns, rather than longer descriptions of the referent. Kurumada and Jaeger (2015) find that the plausibility of grammatical function assignment for two arguments, based on a subjective rating task, is predictive of patterns of optional object marking.

Given these findings that subjective ratings predict behavior for both perception and production, and that both perception and production are sensitive to idiosyncratic

properties of words (see, e.g. Hay & Foulkes, 2016), extending this methodology to the examination of plural predictability influencing /s/ duration is reasonable. The ratings obtained in the perception studies above are either for estimates of frequency of exposure to certain words, estimates of whether certain words are more likely to be used by certain social groups, or estimates of how likely the word *that* is to be used in subordinate clauses. However, the ratings for the production studies are estimating predictability, as in the present study.

The existing studies have found disparate results regarding whether subjective ratings are predictive of behavior. In perception, Walker and Hay (2011) find a significant interaction of word age (calculated from a corpus) and voice age in predicting both error rate and reaction time for a lexical decision task. They do not find this effect when using a subjective rating of whether words are more likely to be used by older or younger speakers.

Kim (2016), however, finds effects of subjective ratings in a lexical decision task in Korean. Rather than comparing corpus ratings and subjective ratings, Kim calculates two scores based on survey data: a stereotype score based on whether participants perceive each word to be used more by older or younger speakers, and a usage-age score, based on self-reported usage of each word, compared across age groups. Then, in a lexical decision task, Kim finds a significant interaction between word age and voice age, in both error rate and reaction time. For error rate, this holds for both the stereotype score and the usage-age score, while for reaction time, it is only significant for the stereotype score.

Kuperman and Van Dyke (2013) also find subjective ratings to be effective predictors of behavior, sometimes even more so than corpus ratings. They use subjective ratings of exposure to different words by participants of varying educational background to predict properties of eye movement and lexical decision latencies in readers of varying proficiency. In this case, subjective ratings are split into two categories – lower education and higher education – which makes them more sensitive. Kuperman and Van Dyke find that their education-sensitive subjective ratings explain either equal or greater amounts of variance in eye tracking and lexical decision behavior compared to corpus-based frequencies. The authors further explore this finding by directly comparing their subjective ratings to the corpus frequencies, and suggest that subjective ratings may actually prove more useful in predicting behavior, because they capture different

sensitivities to frequency at low or high frequencies. For example, the three lowest bins of words (based on corpus frequency) did not show any significant difference in their subjective ratings of frequency within either group of raters, while the high frequency bins all showed robust differences. This suggests that the human raters are more sensitive to frequency differences between higher frequency words than between lower frequency words, which may be why the subjective ratings actually explain more variance than corpus ratings in this study.

In the domain of syntax, Melnick et al. (2010) find that subjective ratings of the likelihood of *that* in relative clauses and complement clauses correlate significantly with predictions of *that* presence based on corpus statistics. These correlations are high for ratings collected both in a lab setting and online via Amazon Mechanical Turk, providing further evidence that online collection of subjective ratings yields comparable results. While this study shows that subjective ratings and corpus statistics are correlated, these measures were not used to predict independent behavior.

In production, several studies have found subjective ratings of predictability to be predictive of the usage of optional linguistic material (e.g. Kurumada & Jaeger, 2015; Kravtchenko, 2014) or of the choice between shorter and longer productions (e.g. Tily & Piantadosi, 2009). As noted above, all of these studies find that when the message (either grammatical function assignment or the identity of the referent) is more predictable, the production is more reduced.

Kuperman and Van Dyke also show that regardless of whether subjective scores are predictive of behavior, they are correlated with corpus-based frequency measures. The present study uses subjective ratings based on both one and five words of context, collected via a crowd-sourcing platform, to predict patterns in production. The one-word ratings can be compared to PWPP ratings, as calculated in Chapter 3, while the comparison between one- and five-word ratings may indicate whether different context sizes are relevant in predicting variation in bound morphemes based on their contextual predictability.

The present study differs from many of the above studies in that while the above studies look at variation across words in terms of the social characteristics of the speakers who produce these words, the present study is examining variation across contexts in which the words are produced. It is an open question whether subjective ratings of a contextual measure like plural predictability are comparable to subjective ratings of the social

characteristics of speakers who commonly produce words. However, this does bear a stronger resemblance to the probability of the word *that* occurring.

4.3 Methods

4.3.1 Experimental setup

The experiment consists of an online rating task carried out on the CrowdFlower platform, where participants are shown a string of words (either one or five words), followed by '_____', and asked to choose whether the string of words would best be followed by *wug* or *wugs*.[37] Nonce words were chosen so as to avoid capturing word predictability, rather than the predictability of plurality. The study in Chapter 3, which showed a significant effect of PWPP, but not of word bigram predictability, suggests that plural predictability is the more important factor in explaining variance in plural /s/ duration.

4.3.1.1 The platform – CrowdFlower

An online platform was chosen because of the ease of recruiting a large number of participants over a short length of time, and because it allows access to a larger body of participants. While the majority of linguistic crowdsourcing research has been conducted via Amazon Mechanical Turk (AMT), as was used in Chapter 2, AMT was not ideal for this project. This was primarily due to the requirement of paying via a US bank account in order to use AMT, but also because the alternative platform used here, CrowdFlower, has a very intuitive and simple way of implementing the type of task required for this project. CrowdFlower has been used for other linguistics experiments (e.g. Wang, Huang, Yao, & Chan, 2014), and works particularly well when the task setup requires repetition of the same type of question many times, with small changes to the content. As all of the questions in this study were identical, only changing the content of the context, this was ideal. First, a csv file was prepared which included all of the contexts, including test questions. Then, after setting up the formatting for one

[37] It is possible that participants have pre-existing ideas about the meaning of wug that may have influenced their judgement. Nonce words were chosen for the reasons stated in the text, and a single pair of nonce words was selected to avoid unforeseen effects of different nonce words. A future study might use the real words from the corpus and compare the results to those found here.

question and uploading the csv, the experiment was ready to run. Questions are automatically randomized in CrowdFlower, and collection for each context stops when the required number of judgments is reached. In addition to the ease of experimental setup, CrowdFlower can be used easily outside the United States, which was an important consideration. However, there are a few limitations, which will be discussed in Section 4.3.3.

4.3.1.2 Instructions

The instructions explain to participants that they will see a string of words, and will then be asked to select which of two non-words best fits after that context. They are told that the non-words represent singular and plural words in English, and that these words could be either concrete or abstract. The full instructions can be found in Appendix 1.

4.3.1.3 Quiz

If participants opt to begin the task, they must first pass a quiz section where their selected responses on a set of ten test questions are compared to a range of optimal responses as selected by the researcher. These test questions are sampled from items which are either highly likely to be plural, or highly likely to be singular. The test questions and the selection of optimal responses are discussed further in Section 4.3.1.5. In order to continue the task, a score of at least 90% on the quiz was required.

4.3.1.4 Task

Once participants have passed the quiz, they are shown pages of ten questions at a time, each of which presents participants with a context and asks them to choose, on a scale of one to ten, whether this context should be followed by a singular or plural word. Singular and plural words are represented by the nonce words *wug* and *wugs*. Participants are able to complete as many ten-question pages as they like. A sample question is shown in Figure 4.1.

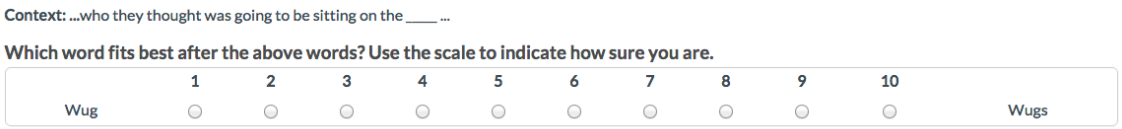

Figure 4.1: Sample question.

While the example above uses ten words of context, the present study used only contexts of one and five words. The contexts, as discussed in Section 4.3.1.6, are extracted from the ONZE corpora. Below is an example of an extracted context:

"…family for instance where there were say four or five kids…"

Given this excerpt, the participant will see one of the following strings of words, depending on the condition:

- "…were say four or five ___ …"
- "…five ___ …".

After selecting a response between one and ten, the participant scrolls down to the next question. At the end of a page of ten questions, responses are recorded and another page is displayed. Participants may stop at any point.

4.3.1.5 Test questions/accuracy

Each page of questions includes one test question (not marked for the participant). If a given participant's accuracy on test questions drops below 90%, that participant is prevented from completing any more questions.

Test questions are questions which are labelled by the experimenter to have one or more 'correct' answers, and are used as a measure of quality control. The test questions for this study were selected from two pools. The first consists of contexts for which the preceding word plural predictability (PWPP) calculated from the corpus is in the top tenth of possible PWPP scores. These preceding words are almost always followed by a plural (e.g. *these, few, several, thousand*). The optimal responses for these test questions were set to eight, nine, and ten. The second pool consists of contexts extracted from before singular nouns, ending in words which are almost always followed by a singular noun (e.g. *a, that*). The optimal responses for these test questions were set to one, two, and three. A range of options for optimal responses was given in order to allow for variation in participant use of the scale.

4.3.1.6 Extracted contexts

The contexts which are shown to participants are excerpts from three collections of the ONZE corpora, as detailed in Chapter 3 (Section 3.3.2), and consist of either one or five words. For each plural word used in the corpus study from Chapter 3, the preceding context from the corpora was extracted. For the purposes of this experiment, only the transcriptions of these recordings are used, not the recordings themselves. A subset of the contexts used can be found in Appendices 4 and 5.

Steps were taken to ensure that all of the excerpts used in this experiment are anonymous, and that they do not contain potentially offensive material. All content deemed inappropriate was excluded (e.g. referring to drugs, sexuality, or violence), as well as any references to Māori culture. All names were changed, except those of public figures. In total, there were 776 unique one-word contexts (compared to 789 used in Chapter 3), and 4,938 unique five-word contexts (compared to 5,047 in Chapter 3) that were tested. This corresponds to 5,253 total tokens that were coded for one-word contextual predictability (99.6% of those used in Chapter 3), and 4,994 total tokens that were coded for five-word contextual predictability (94.7% of those used in Chapter 3).

4.3.1.7 Participants

Participants were recruited through CrowdFlower, where they select a task based on the title ("Choosing singular or plural words based on context"). After opening the task, participants read the instructions and choose whether or not to complete the task.

The participants are speakers of English who reside in the United States and who are registered workers at CrowdFlower, the online platform used to run the experiment (see Section 4.3.1.1). Speakers in the United States were chosen because there is a large body of workers who are located in the United States, and as they all have a similar grammar of English, they should have similar judgments about how predictable a plural is, given a preceding context. Although ideally, speakers of New Zealand English would have been recruited, there is not a large enough body of New Zealand participants enrolled in any online platform to satisfy the requirements of this study. Nonetheless, there is some evidence that participants from different dialect regions perform similarly on rating tasks with stimuli from one of these dialect regions, as seen in Walker and Hay (ms). That study finds effects of the congruence between word gender and speaker gender facilitating lexical decision using New Zealand English stimuli, for both New Zealand and United States participants. This suggests that NZ and US participants have similar statistical representations of how often certain words are used by men and women. While plural predictability is quite different from word gender, the similarity in one suggests a reasonable amount of crossover between the statistical properties of the two dialects.

4.3.1.8 Information sheet and consent

Because CrowdFlower does not allow for a consent form, and all workers on CrowdFlower have already given consent for their work to be used, there was no

separate consent form. However, in the instructions, participants were provided with a link to an information sheet detailing the purpose of the project and giving them the opportunity to contact the researcher with any questions. The information sheet is included in Appendix 2.

4.3.2 Pilot to determine the number of judgments necessary

Before running the entire set of contexts, a sample set was run in order to determine how many judgments were needed per item. If mean scores based on five judgments per item are closely correlated with mean scores based on 20 judgments, then using only five judgments per item can be justified.[38] In the initial run, 330 items were tested. These came from 110 different items from the corpus, using either one, five, or ten words of preceding context for each item. These 110 items were sampled from across the range of PWPP scores calculated from the corpus (based on one word of context), with ten items in each decile of the PWPP range, as well as ten items which were extracted from before singular nouns in the corpus. The items preceding singular nouns were selected such that they should be unambiguously followed by a singular noun (e.g. ending in the word *a, one, each*). Each context of one, five, or ten words was rated by 20 participants. The ten singular contexts and the ten plural contexts from the highest decile were tagged as test questions, which allowed for monitoring of how well the participants performed. If performance on the test questions fell below 90%, participants were forced to stop the task and their data was discarded.

In order to determine how many judgments were necessary per item, sets of five or ten judgments for each item were randomly sampled from the 20 judgments collected, in order to simulate collecting fewer judgments. The mean score for each context was then calculated based on five, ten, or 20 judgments, and correlations were calculated between the scores from five and 20 judgments, as well as between ten and 20 judgments. This procedure was repeated 200 times. While correlations for ten and 20 judgments were always higher than between five and 20, the lowest correlation parameter found was .87, between five and 20 judgments of five-word contexts. Figure 4.2 shows histograms of

[38] Note that the high correlation between the means does not take into account the amount of variability across the ratings. However, correlations between means remained high through 200 repetitions of this procedure.

the correlation parameters for the 200 samples for each pairing of number of judgments and context size. Based on these high correlations, a decision was made to use five judgments per item. In the end, only one- and five-word contexts were used, but this pilot study demonstrated that for future work, five judgments is also acceptable for ten-word contexts.

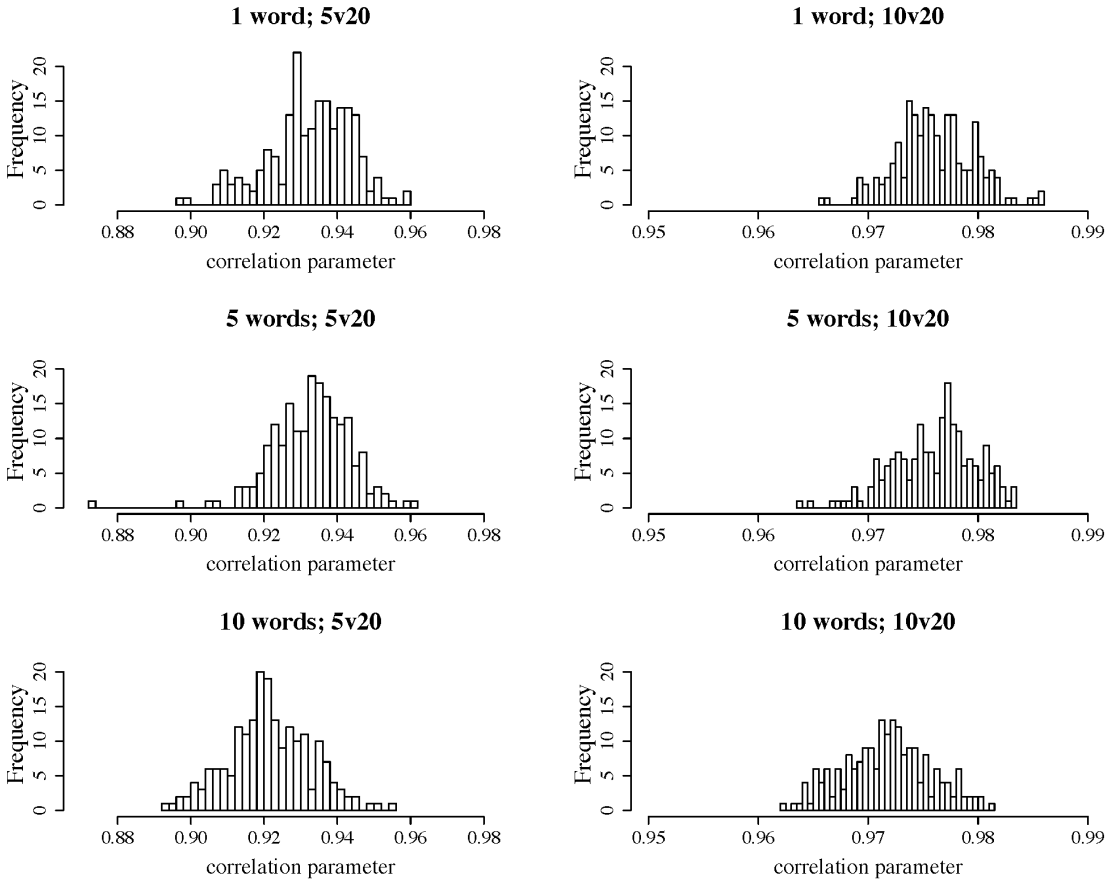

Figure 4.2: Histograms of correlation parameters across 200 trials.

4.3.3 Using CrowdFlower

As discussed in Section 4.3.1.1, the crowdsourcing platform CrowdFlower was used to carry out the rating task. While there were advantages to using this platform, as discussed above, there were also several drawbacks to using CrowdFlower. This may mean that it would be better to use a different platform in future studies.

The most substantial problem is that it was not possible to collect demographic information such as age, gender, education level, or linguistic background from the participants, as CrowdFlower does not permit such questions. The participants were limited to speakers of English with an IP address in the United States, who were certified at Level 2 of 3 at the minimum, indicating moderate experience and accuracy.

Additionally, while it was possible to ensure that no participant rated a given item more than once, there was not at the time a way of limiting how many responses could be collected from each participant. This means that a single participant could potentially respond to all 5000 items, while another might only respond to ten. The following section outlines some characteristics of the individual variation in ratings.

4.3.3.1 Individual variation in ratings

As mentioned above, there was no limit on how many responses could be collected from a given participant. Figure 4.3 shows the distribution of the number of participants who completed different numbers of ratings. For the five-word contexts, there were 58 participants, ranging in the number of contexts completed from 10 to 4,134 (excluding singular contexts). For the one-word contexts, there were 68 participants, and a range of 10 to 153 plural contexts rated.

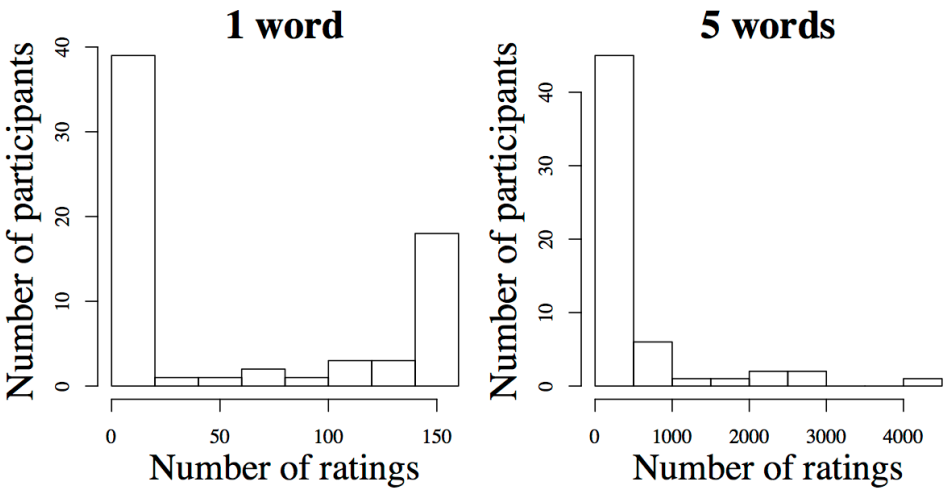

Figure 4.3: Histograms of number of contexts rated per participant.

The response patterns of individuals varied substantially. Some used the whole scale, while some used only 1 and 10. Due to this variability, several different analyses were performed. Figure 4.4 shows a few sample distributions:

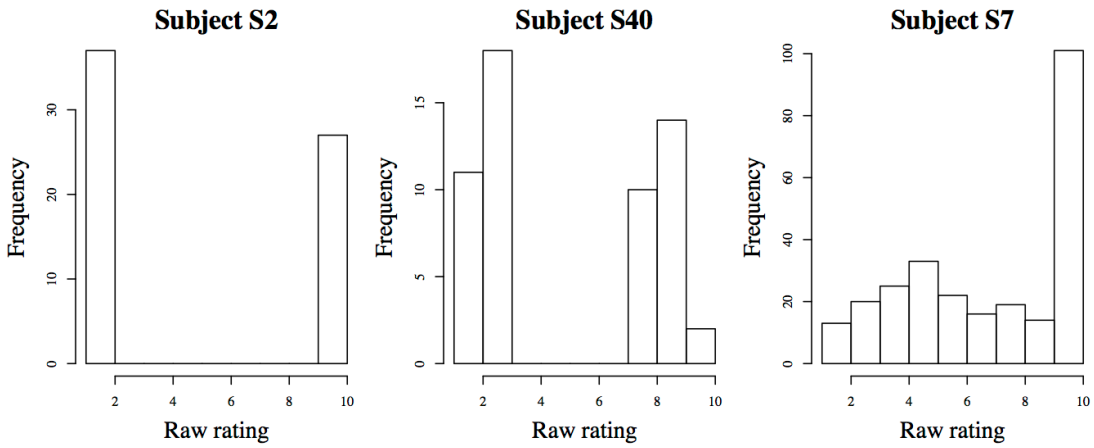

Figure 4.4: Sample distributions of ratings by participant.

4.3.4 Analysis

The analysis proceeds in several steps, starting with a basic calculation of plural predictability, then adding in levels of complexity to account for variation in the ratings. At each step, a preceding context plural predictability (PCPP) score is calculated for each item (based on the ratings of either a 1-word or 5-word preceding context). This score is then tested across all items in the dataset for correlation with both the preceding word plural predictability (PWPP; calculated in Chapter 3) and other PCPP scores. Finally, for each step, these PCPP scores are tested to see whether they are predictive of plural duration in the dataset. This is tested by beginning with the final model presented in Chapter 3 and adding the new PCPP score, both with and without PWPP in the model. Each PCPP score is also tested to see whether it interacts significantly with any of the other factors. If a PCPP score significantly improves the model fit, either alone or in an interaction with another factor, it is retained. This is determined by conducting ANOVA tests between models with and without the given PCPP score, and comparing Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) scores. If the ANOVA test shows a significant ($p < .05$) improvement in model fit, the factor is retained. The model from Chapter 3, used as a base for this testing, is shown below:

```
lmer (log plural duration ~ PWPP + onset coronal obstruent +
      coda gradient similarity score * log speech rate + log speech rate +
      log base duration + plural type + preceding phone manner +
      base onset complexity + base coda complexity + corpus + vowel length +
      next phone manner +  preceding word bigram predictability +
      (1 + PWPP | speaker) + (1 + PWPP | plural word))
```

The different ways in which PCPP scores are calculated are, for both one-word and five-word preceding contexts: mean raw ratings, mean z-scored ratings, and mean binary ratings, as well as several types of extracted by-context intercepts. These by-context intercepts are further discussed in Section 4.3.4.4. The remainder of this section outlines each way in which PCPP scores are calculated and the reasons for each method.

4.3.4.1 Mean raw ratings

For each preceding context, five participants gave a rating on a scale of one to ten, reporting whether the context is likely to be followed by a plural or singular (10 = plural; 1 = singular). The first PCPP score used is simply a mean of the five ratings for each context. While this measure does not take into account any individual variation in the way participants rate contexts, if there is a strong effect of speaker intuition about morphological predictability, this measure will show it. The distribution, not including singular items, of PCPP scores calculated as mean raw rating is shown below (1 word: range 1 – 10, mean 5.84; 5 words: range 1.2 – 10, mean 6.84):

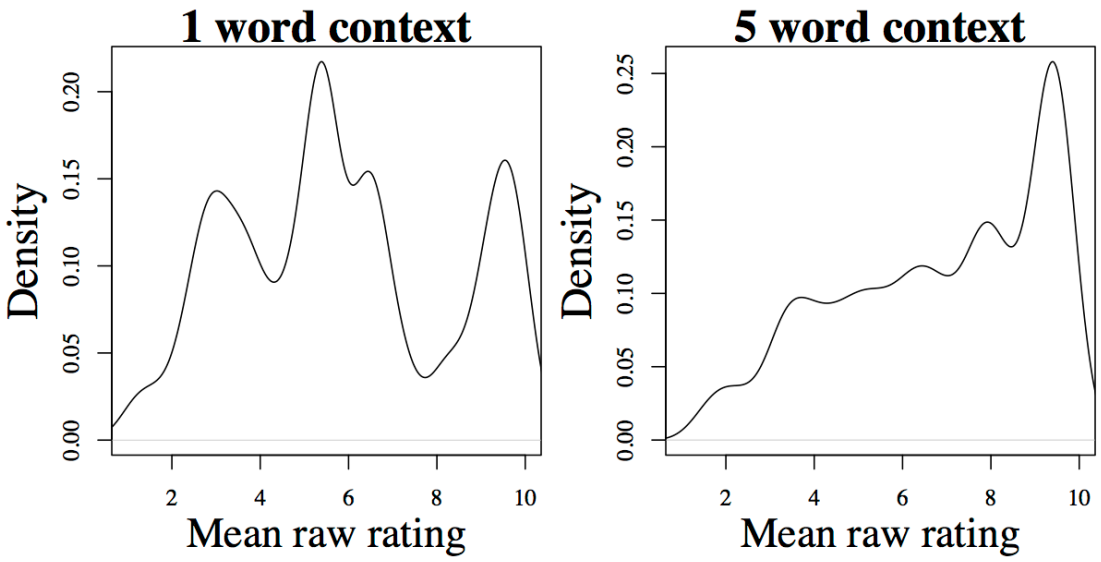

Figure 4.5: Distribution of PCPP scores based on mean raw ratings.

4.3.4.2 Mean z-scored ratings

As discussed in Section 4.3.3.1, the behavior of individual participants in rating contexts was not uniform. In order to account for individual differences in using the scale, another PCPP score was calculated. For this score, ratings by each participant were first z-scored, then the mean of the five z-scored responses for each context was calculated. The distribution of PCPP scores calculated as mean z-scored rating is shown below (1 word: range -1.29 – 2.82, mean 0.32; 5 words: range -1.71 – 1.24, mean 0.10):

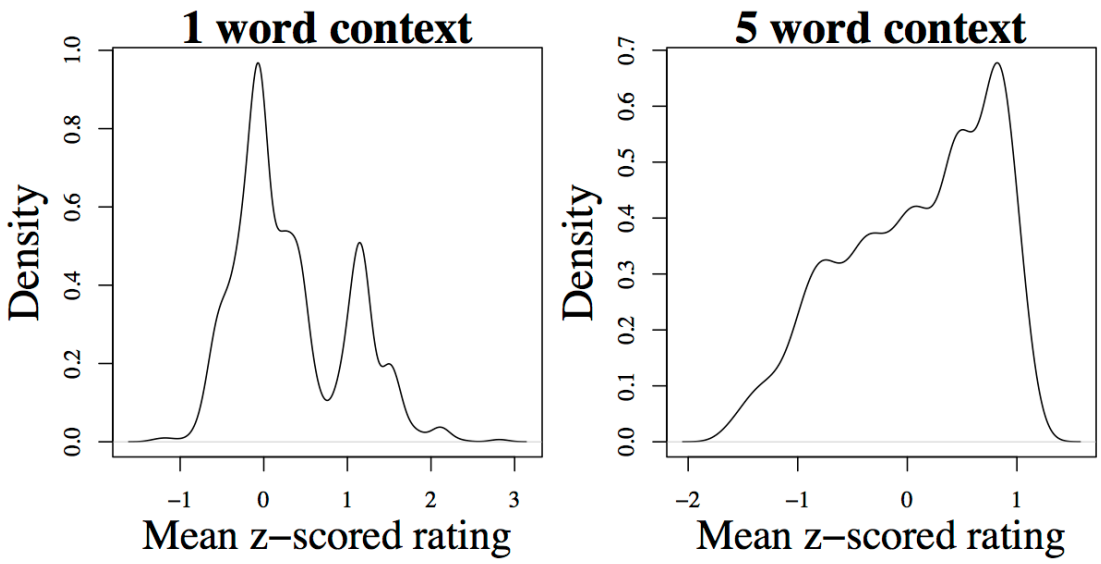

Figure 4.6: Distribution of PCPP scores based on mean z-scored ratings.

4.3.4.3 Mean binary ratings

In addition to differences in the range of numbers used in the scale, many participants used only one or two numbers, making their ratings essentially binary. A third PCPP score was calculated by finding the mean score for each participant, then classifying each rating as either likely to be singular (0) or likely to be plural (1), based on whether the rating was above or below the mean rating for that participant. The mean of the five binary scores was then taken for each context. The distribution of PCPP scores calculated as mean binary rating is shown below (1 word: range 0 – 1, mean 0.58; 5 words: range 0 – 1, mean 0.63):

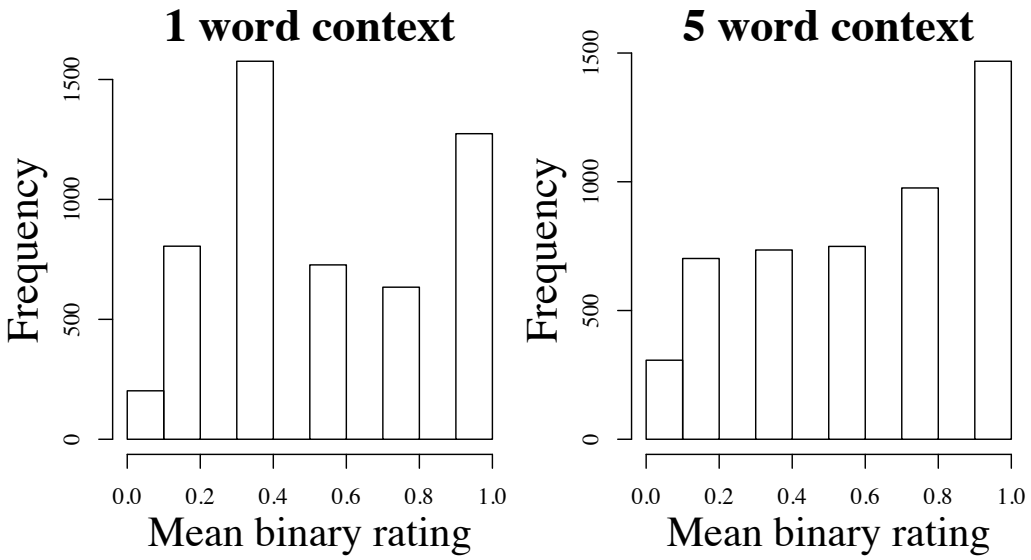

Figure 4.7: Distribution of PCPP scores based on mean binary ratings.

One additional score based on binary ratings was calculated, an overall singular or plural binary score based on the mean binary score.

4.3.4.4 Intercepts for each context

In addition to differences in how participants used the scale, there may have been effects of task adaptation on how participants rated contexts. For example, participants may have gotten more comfortable with the scale as the task progressed, or may have gotten fatigued and reduced the amount of variation in their responses. Recall that the measure of progress through the task that is available from CrowdFlower is page number, rather than item number. In order to account for potential progress effects, three more PCPP scores were calculated. For each of these three scores, a linear mixed effects model was used to predict the rating (either raw, z-scored, or binary) as a function of page number, with a random intercept by context, a random intercept by participant, and a random slope of page number by participant. However, the random slope of page number by participant was only found to significantly improve model fit for one of the six models (5 words, binary), so it was eliminated from all the others. In each model, the fixed effect of page number was tested as both a linear and non-linear effect (by allowing the effect to be a second-degree polynomial). However, the non-linear component was never found to significantly improve model fit. One of the six models used to calculate by-item intercepts is shown below. All other models followed this pattern, with the exception of the 5-word binary model, which included a random slope of page number by participant.

```
lmer (raw rating one word ~ page number + (1 | participant) +
      (1 | one word context))
```

In this model, the ratings are explained to the degree possible by progress through the task (the fixed effect of page number), as well as by individual variation in how each participant used the scale (random intercept by participant) and differences across contexts (random intercept by context). The random intercept by context captures an estimate of the rating for each context, taking into account effects of page number and participant. This is essentially a mean rating for each context after accounting for progress through the task and by-participant variation. This by-context random intercept was extracted and used as another PCPP score. The process was repeated three times for each of the two context sizes, using raw rating, z-scored rating, and binary rating as the outcome variable in the linear mixed effects model. The distributions of PCPP scores calculated in this way are shown below:

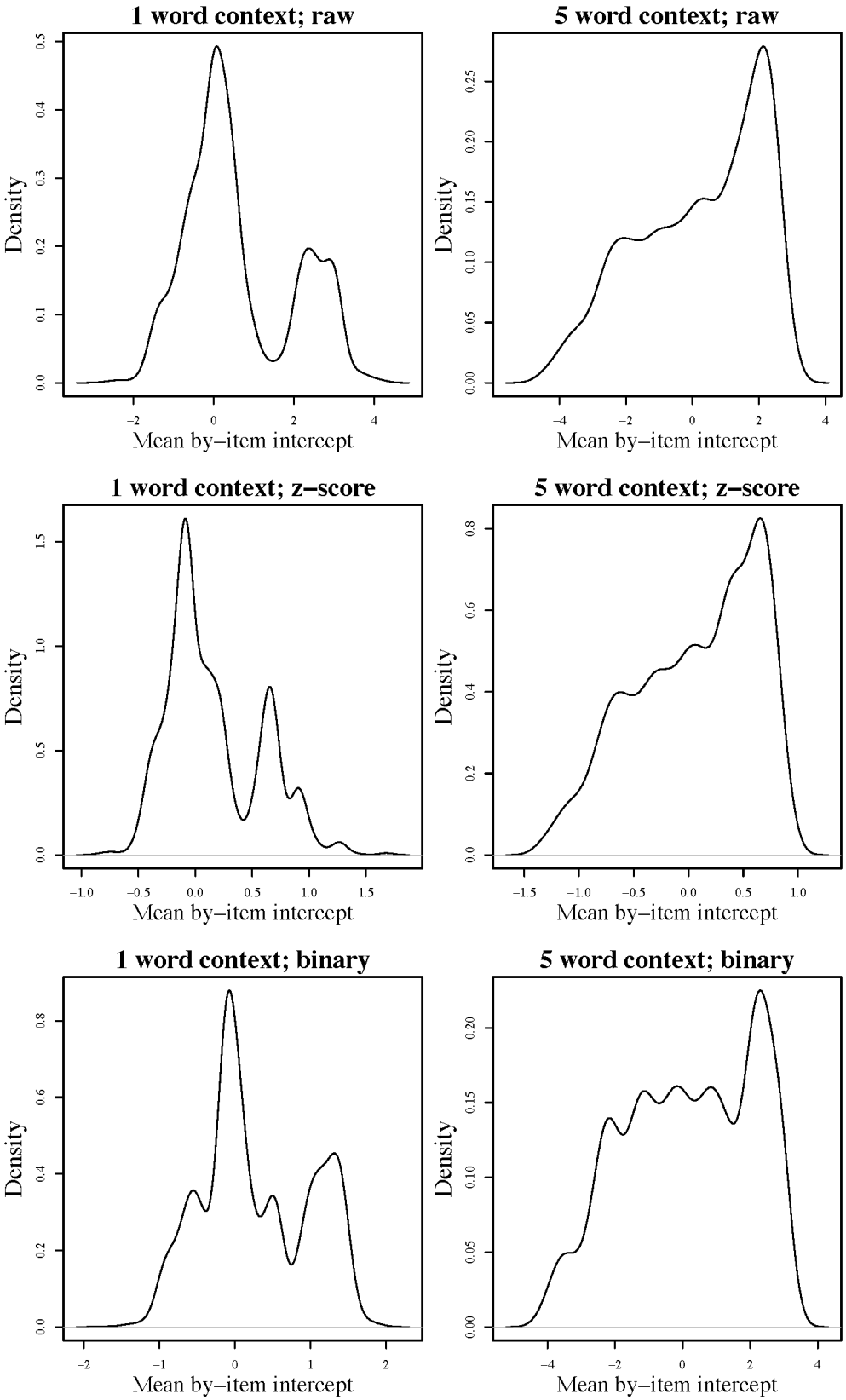

Figure 4.8: Distributions of PCPP scores based on by-item intercepts.

Table 4.1: Range and Means of PCPP Scores Based on By-Item Intercepts.

PCPP (by-item intercepts)	min	max	mean
1 word raw ratings	-2.76	4.22	0.59
1 word z-scored ratings	-0.84	1.67	0.16
1 word binary ratings	-1.74	1.95	0.28
5 word raw ratings	-4.69	3.17	0.23
5 word z-scored ratings	-1.39	1.00	0.07
5 word binary ratings	-4.29	3.38	0.18

4.4 Results

4.4.1 Correlations

In total, twelve preceding context plural predictability (PCPP) scores were calculated: six for one-word contexts and six for five-word contexts. For both one-word and five-word contexts, there were three scores calculated by taking the mean across five ratings (raw, z-scored, binary), and three scores calculated by extracting the random intercepts by context from a linear mixed effect model predicting ratings (with raw, z-scored, or binary ratings used as the outcome variable). Table 4.2 shows the Spearman rank correlations of each of these PCPP scores with the preceding word plural predictability (PWPP) score calculated in Chapter 3, as well as each of the other PCPP scores. Also included is the PWPP score calculated only across following nouns ($PWPP_{noun}$), rather than all words. This score may be closer to the PCPP scores, as participants were only given a choice between plural or singular nouns, rather than any word. In the table, green cells show the highest correlations, while yellow cells are the lowest. All correlations are significant ($p < .01$), as tested using the rcorr function in R (Harrell 2016).

Within the PCPP scores based on one-word contexts, as well as within those based on five-word contexts, the correlations are very high, all at or above 0.83. This is not too surprising, given that they are all based on the same ratings, but it does indicate that the effects of by-participant variation and progress through the task are minimal. Between the one-word and five-word PCPP scores, correlations are still moderate, but not as high, ranging from 0.54 to 0.58.

Table 4.2: Spearman Rank Correlations of All PCPP Scores

	logPWPP	logPWPP - noun	Mean rating (raw), 1word	Mean rating (z-scored), 1 word	Mean rating (binary), 1 word	Intercepts (raw), 1 word	Intercepts (z-scored), 1 word	Intercepts (binary), 1 word	Mean rating (raw), 5 words	Mean rating (z-scored), 5 words	Mean rating (binary), 5 words	Intercepts (raw), 5 words	Intercepts (z-scored), 5 words	Intercepts (binary), 5 words
logPWPP	X													
logPWPP - noun	0.96	X												
Mean rating (raw), 1word	0.51	0.54	X											
Mean rating (z-scored), 1 word	0.49	0.54	0.89	X										
Mean rating (binary), 1 word	0.48	0.52	0.91	0.92	X									
Intercepts (raw), 1 word	0.49	0.54	0.91	0.98	0.92	X								
Intercepts (z-scored), 1 word	0.48	0.53	0.89	1	0.92	0.98	X							
Intercepts (binary), 1 word	0.46	0.51	0.83	0.94	0.92	0.94	0.94	X						
Mean rating (raw), 5 words	0.49	0.51	0.56	0.57	0.57	0.56	0.57	0.57	X					
Mean rating (z-scored), 5 words	0.5	0.52	0.56	0.58	0.58	0.56	0.58	0.57	0.96	X				
Mean rating (binary), 5 words	0.48	0.5	0.54	0.56	0.56	0.54	0.56	0.55	0.97	0.97	X			
Intercepts (raw), 5 words	0.5	0.52	0.57	0.58	0.58	0.57	0.58	0.58	0.98	0.99	0.97	X		
Intercepts (z-scored), 5 words	0.5	0.52	0.56	0.58	0.58	0.56	0.58	0.58	0.97	1	0.97	1	X	
Intercepts (binary), 5 words	0.49	0.51	0.54	0.56	0.56	0.55	0.56	0.56	0.94	0.98	0.97	0.97	0.98	X

Between PWPP/$PWPP_{noun}$ and the PCPP scores, correlations are lower, ranging from 0.48 to 0.54. However, these correlations are all significant. One additional observation can be made about the correlations, specifically with respect to the difference between PWPP and $PWPP_{noun}$. Across the board the correlations with $PWPP_{noun}$, a measure which more closely approximates the task of choosing between a singular and plural noun, are higher than those with PWPP. This indicates that the PCPP scores are capturing, to some extent, the probability of a plural noun following the given context. This is encouraging, because it indicates that the task was capturing something related to plural predictability, even if it was not as nuanced as expected. However, given that correlations with corpus measures were only moderate, it is unclear whether the effect from Chapter 3 will be replicated with the subjective measures.

4.4.2 PCPP scores and /s/ duration

After running models with each of the different PCPP scores, none of the scores emerged as a significant predictor of /s/ duration in the corpus, either alone or in interactions. The PCPP score which had the largest t-value (-.707) was the mean binary score for the five-word contexts. This effect is in the expected direction, but far from significant, meaning that this experiment does not provide evidence that subjective ratings of the likelihood of plurality based on one or five words of preceding context are predictive of variation in plural /s/ duration.

4.5 Discussion

4.5.1 Possible reasons for null result

There are three possible reasons for the null result presented in Section 4.4.2: (1) the larger context is not relevant in predicting plural duration; (2) human-based ratings of plural likelihood are not as relevant to predicting plural duration as probabilities from a corpus; (3) this task did not capture the relevant statistical knowledge that language users have about plural probabilities given context.

The first possibility, that the larger context is not relevant, is certainly possible. However, if it was a problem of context size, then the one-word PCPP ratings should have been predictive, as the PWPP score calculated from the corpus was. The observation that corpus-based probabilities given one word were predictive, while subjective scores were not, suggests that this study does not exclude the possibility of the larger context being relevant to predicting plural /s/ duration.

The second possibility, that human-based ratings are not relevant, is perhaps more likely. However, it is possible that the problem lies in the task that was used to collect these human-based ratings rather than in the concept of using human-based ratings at all. This study shows that overt ratings of plural predictability are not predictive of plural duration, but there may be other human-based scores which are. It is possible that the task used in this study captures conscious knowledge, while something as subtle as the predictability of a plural given preceding context is something that is not available at a conscious level. If this is the case, then a different task which captures subconscious knowledge would be more appropriate, such as a study using eye tracking or measuring reaction times. This result provides no evidence to support Hypothesis 5:

> **H 5:** Language users have conscious access to knowledge of the statistical properties of morphemes.

This result does not provide conclusive evidence regarding Hypothesis 6:

> **H 6:** More than one word of context is used to track the predictability of morphemes.

4.5.2 Future directions

Given the null result presented in this chapter, one future direction would be to create another experiment which captures subconscious knowledge about plural probabilities. In the domain of language change and attitude, there is evidence that explicit and

implicit attitude ratings tend to differ, and that it is the implicit attitudes which are predictive of language change, rather than explicit attitudes (Kristiansen, 2009, 2010). It is possible that ratings of plurality are similar, in that while overt, conscious ratings of how likely a plural is are not predictive of /s/ duration, there may be some implicit knowledge that this task did not capture, which could be predictive. A task measuring reaction time, or using eye tracking, would be more likely to capture such implicit knowledge. One such task would be a yes/no acceptability judgement, using the same contexts and presenting each context with both the singular and plural nonce word, while measuring reaction time. In order to have the maximum chance of success, this task should be run with speakers of New Zealand English because their grammar matches most closely with that of the speakers in the corpus.

The task could also be conducted using the real words used in the corpus, rather than nonce words, in order to make it closer to language experienced by participants in the real world. However, this would allow for the possibility of wordform conditional probability to drive the ratings, rather than plural predictability. Consequently, it might be useful to conduct two studies, one using nonce words and one using real words, to see if one type of rating proves more predictive than the other. Yet another option would be to give an open response after the selected context, then code the responses for whether they are plural, using this measure to calculate plural predictability.

If future studies are able to capture subconscious ratings of plural predictability using one or five words of context, they should explore the effects of using larger context sizes, to determine whether there is a limit on how much context is relevant for predicting variation in the production of bound morphemes.

Finally, it would be interesting to calculate measures of plural predictability based on larger contexts, using a large corpus such as the Web 1T 5-gram corpus (Brants & Franz, 2006), and comparing these scores to both the one word score from the ONZE corpora, and the multi-word scores calculated from other human-based ratings. However, as discussed in Section 4.1, plural predictability scores from a corpus would not capture plural predictability in the same way as human-based ratings. If corpus-based ratings of larger contexts were also predictive of plural /s/ duration, it would indicate that it is the specific sequence of words that is relevant, rather than the more abstract idea of plurality captured by human ratings, as expected in this study.

4.5.3 Conclusions

Although this task yielded a null result, it opens up the door for further exploration of how much context is relevant when calculating the predictability of bound morphemes. In the MOP framework, as in other Bayesian approaches to linguistics, while context is shown to be important, the exact nature and size of the context that is important to which research questions is an open question. In order to continue informing this vein of research, a follow-up study which uses a different methodology to try to access language users' statistical knowledge of plural predictability over different context sizes would be very useful. While explicit subjective ratings were not predictive of plural /s/ duration, either with one word or five words of context, it may be that implicit ratings would be predictive. Additionally, using a large web corpus to estimate predictability based on a variety of context sizes may be informative.

5 OVERALL DISCUSSION

5.1 Summary of research questions, hypotheses, and findings

This thesis uses an artificial language learning experiment and a corpus of New Zealand English to show that both the learning and production of morphological cues are influenced by the amount of information those cues carry. Morphological cues which are more predictable, and therefore carry less information, are learned more poorly and produced with more reduced realizations. Using the plural as a case study, these findings show that language users track the statistical properties of bound morphemes, and provide further evidence that language users have knowledge of the statistical properties of these morphemes, independent of the words to which they are attached. In turn, this suggests that language users have knowledge of the predictability of morphological cues, contributing to the understanding of what constitutes a language user's knowledge of language.

As discussed in Chapter 1, this research builds on work showing the influence of contextual predictability on linguistic behavior at many levels of linguistic structure (e.g. Aylett & Turk, 2006; Bell et al., 2003; Cohen Priva, 2008, 2012, 2015; Jaeger, 2010, 2011; Jurafsky et al., 1998, 2001, 2002; Raymond et al., 2006; van Son & Pols, 2003; for review, see Jaeger & Buz, 2017). While some previous work has examined the influence of contextual morphological predictability on morpheme reduction or omission (e.g. Bybee & Scheibman, 1999; Cohen, 2014, 2015; Frank & Jaeger, 2008; Kurumada & Jaeger, 2015; Norcliffe & Jaeger, 2016; Seyfarth, 2016), the amount of work in this area is limited, and has been conducted almost exclusively in an

experimental setting. This thesis builds on that work by first testing whether language users are sensitive to the statistical usage patterns of morphological cues to plurality in an Artificial Language Learning task, then examining production behavior in a corpus of naturally occurring speech. Using the plural as a case study, the findings from this thesis demonstrate that, as at other levels of linguistic structure, predictability influences linguistic behavior with regard to morphemes.

The proposed motivation for these influences of predictability on behavior is that language is a system of message transmission, subject to the biases of successfully transmitting messages and conserving resource cost (as discussed in Hall et al., submitted). Linguistic units which are more predictable carry less *information* (Shannon, 1948), which means they are less important to successful message transmission. This makes them good targets for reduction, resulting in patterns where less predictable units are learned quickly or produced with less reduced realizations, while more predictable units are learned later and reduced in production. In this thesis, the amount of information carried by a morphological cue is calculated based on the predictability of the message of plurality, given the context. In order to assess the influence of predictability on morphological behavior, two frameworks which approach linguistic behavior in terms of information are used: the Rescorla-Wagner (RW) model (Rescorla & Wagner, 1972) and Message-Oriented Phonology (MOP; Hall et al., submitted). While the RW model does not explicitly state that language is a system of information transfer, it does frame the learning of cues in terms of how predictive they are of the outcome (message) in combination with other available cues. It predicts that cues which contribute more to identifying the message will be learned better. Message-Oriented Phonology explicitly discusses language in terms of information, outlining communicative biases that shape any effective communication system. MOP emphasizes that language is meant to transmit *messages*, rather than individual linguistic units such as phonemes.

Using these two frameworks to quantify the amount of information carried by morphemes assumes that morphemes have, at least to some extent, representations which are independent of the words to which they are attached. While there are arguments both for and against the idea that morphemes have independent representations (e.g. Hanique & Ernestus, 2012; Hay, 2004; Plag et al., 2017; Schuppler et al., 2012; Zimmermann, 2016), the findings of this thesis contribute to the evidence in favor of independent representations.

The three studies presented in this thesis address the research questions and hypotheses presented in Chapter 1, using the two frameworks presented above. The findings confirm Hypotheses 1-3 (as will be reviewed below), demonstrating that language users do have knowledge of the statistical properties of morphemes. This is done by providing evidence that linguistic cues signaling plurality are learned less well and are produced with more reduced realizations when they are more predictable. The findings also provide evidence in favor of Hypothesis 4, that bound morphemes have independent representations of some kind. In addition to these confirmed hypotheses, the findings provide partial evidence against Hypothesis 5, that language users have conscious access to knowledge of the statistical properties of morphemes. Finally, the findings are inconclusive with regard to Hypothesis 6, which states that more than one word of context is used to track the predictability of morphemes.

Study 1, an online artificial language learning experiment, demonstrates that language learners use morphological cues less when the message they are signaling is more predictable. Learners who were exposed to a language in which the message of plurality was entirely predictable without Cue B used that cue less than learners who were exposed to a language in which the message of plurality was not always predictable without Cue B. Study 2 demonstrates that the production of English plural /s/ is sensitive to the morphological predictability of plurality, with more predictable plurals having shorter /s/ duration. Finally, Study 3 shows that, while subjective ratings of plural predictability are correlated with corpus-based measures, these subjective ratings are not predictive of plural /s/ duration. Neither the one-word nor five-word measures of subjective plural predictability were found to be predictive.

The research questions and hypotheses presented in Chapter 1 are presented again below, with the evidence either supporting or refuting them.

RQ 1: What constitutes a language user's knowledge with respect to morphemes?
RQ 1a: Do language users have knowledge of the predictability of morphological cues?
H 1a: Language users do have knowledge of the statistical properties of morphemes.

Hypothesis 1a is motivated by previous research showing that language users have knowledge of the statistical properties of linguistic units at a variety of levels of linguistic structure. Hypothesis 1a is supported by the findings in both Study 1 and Study 2. Study 1 shows an influence of the predictability of morphological cues on the

learning of those cues in an artificial language, while Study 2 shows an influence of the predictability of the NZE plural /s/ on the production of that /s/. Both of these results imply that language users do have access to details about the statistical properties of morphemes.

RQ 2: How is the learning of linguistic cues which signal the grammatical category of plurality influenced by predictability?
H 2: Linguistic cues signaling plurality are learned less well when the message they signal is more predictable.

Hypothesis 2 is derived from previous work showing that linguistic units which are more predictable tend to be learned later or with more difficulty (e.g. Dietrich et al., 1995; Ellis, 2006a, 2006b; MacWhinney, 1997; Ramscar et al., 2013; Ramscar et al., 2013; for review, see Bardovi-Harlig, 1999), and is supported by the findings of Study 1 (Chapter 2). Cue B was learned less well when participants were exposed to a language in which the message it signaled was predictable.

RQ 3: How is the production of linguistic cues which signal the grammatical category of plurality influenced by predictability?
H 3: Linguistic cues signaling plurality are produced with more reduced realizations when they are more predictable.

Hypothesis 3 is derived from previous work showing that linguistic units which are more predictable tend to be produced with reduced realizations (e.g. Aylett & Turk, 2006; Bell et al., 2003; Cohen Priva, 2008, 2012, 2015; Jaeger, 2010, 2011; Jurafsky et al., 1998, 2001, 2002; Raymond et al., 2006; van Son & Pols, 2003; for review, see Jaeger & Buz, 2017), and is supported by the results of Study 2 (Chapter 3), in which NZE plural /s/ durations are found to be shorter when plurality is more predictable. The predictability of plurality is measured here through Preceding Word Plural Probability (PWPP), which is a measure of how likely a plural is to occur, given the word immediately preceding the plural.

RQ 4: Do bound morphemes have some degree of representation that is independent of the words to which they are bound?
H 4: Bound morphemes do have independent representations of some nature.

Hypothesis 4 is motivated by the ongoing debate about whether or not the representation of morphemes is independent to some extent from representations of whole words. Hypothesis 4 is also supported by the findings of Study 2. Study 2 shows that /s/ duration is influenced by the predictability of plurality, while controlling for both wordform frequency and word bigram predictability. This implies that the statistics of morphemes are being tracked independently of word-level statistics. In turn, this

suggests that morphemes do have independent representations of some kind. It is important to note, however, that this study uses the plural, which is among the most productive and most decomposable morphemes. While it is likely that these findings will extend to other bound morphemes, it may be that the statistical properties of less productive or less decomposable morphemes are less likely to be tracked independently of the words in which they occur.

RQ 5: Is this knowledge of statistical properties of morphological cues available at a conscious level?
H 5: Language users have conscious access to knowledge of the statistical properties of morphemes.

Hypothesis 5 is based on previous studies showing a correlation of subjective, conscious ratings with corpus-based measures in studies related to word frequency (e.g. Kuperman & Van Dyke, 2013), and whether certain words are used more by certain social groups (Kim, 2016; Walker & Hay, 2011), as well as previous studies showing a correlation between subjective ratings of predictability and linguistic behavior (e.g. Kurumada & Jaeger, 2015; Kravtchenko, 2014; Tily & Piantadosi, 2009). This hypothesis was tested in Study 3 (Chapter 4) by collecting subjective ratings of plural predictability (Preceding Context Plural Probability, PCPP), then both comparing them with the corpus-based measure (PWPP) and using the PCPP scores as predictors of plural duration. While the one-word PCPP score is moderately correlated with PWPP, it is not a significant predictor of plural duration. This suggests that, while language users have some conscious knowledge of plural predictability, it is either not as nuanced as the corpus measure, or this task was not the appropriate task to capture that nuanced knowledge. This provides some evidence against Hypothesis 5.

RQ 6: What is the size of the context used to track the predictability of morphemes?
H 6: More than one word of context is used to track the predictability of morphemes.

Hypothesis 6 is based on the idea that information signaling plurality may occur more than one word before the plural (e.g. ***A few*** *incredibly cute dog****s***). Therefore, it is likely that a context including more than one preceding word is more predictive than a context of only one word. This hypothesis is also tested in Study 3, and is neither confirmed nor denied, as neither the one-word or five-word subjective score is predictive of plural /s/ duration. The proposed reason for this null result is that this task did not capture fine-grained intuition about plural predictability for either context size. Alternative studies

are proposed in Chapter 4 which would allow for comparison of plural predictability based on contexts of different sizes.

Together, the findings from these studies show that linguistic behavior related to morphological cues is influenced by the predictability of the message of plurality, and suggests that morphemes have some kind of independent representation. These findings provide further evidence that language users have access to statistical knowledge about the usage patterns of morphemes. While previous work has shown the influence of statistical properties on behavior at other levels of linguistic structure, and some work has addressed this question at the level of the morpheme, this thesis provides further evidence suggesting that biases related to effective communication are active at the level of the morpheme, in both artificial language learning and naturally-occurring speech.

5.2 Limitations and future directions

The studies presented in this thesis provide evidence that language users have knowledge of the statistical properties of bound morphemes. However, only one morpheme is used as a case study. Future studies should examine how the learning and production of other morphemes, both inflectional and derivational, are influenced by predictability. Extending this work to other morphemes would provide further information about whether the tracking of statistical properties is a general property of morphemes, or if it only applies to certain morphemes. Hay (2004) proposes that complex words form a continuum from highly decomposable to less decomposable. While plural /s/ is semantically transparent and highly productive, and therefore most words which contain plural /s/ are highly decomposable, this is not true of all morphemes. It may be that language users tend to track the statistical properties of inflectional morphemes because they are transparent and productive, but that this is not true of derivational morphemes. Or, inflectional and derivational morphemes that tend to be transparent may be tracked independently, while others are not. Further exploration regarding how a variety of morphemes behave would provide further information concerning how information about morphemes is stored and accessed.

Future studies should also expand this work to other languages. While evidence of tracking the predictability of plural morphemes is found in both an Artificial Language and New Zealand English, the participants for the ALL experiment were also speakers of English. Examining equivalent phenomena in other languages would provide further

evidence that the influence of predictability on morphological cues is a property of communication systems, rather than any one language in particular.

In addition to expanding to other morphological categories and other languages, future studies could make use of alternative methodologies to explore the question of how large the relevant context is. Study 3 uses an overt rating task to solicit ratings of plural predictability, and finds that, while these ratings are correlated with corpus-based ratings, they are not predictive of /s/ duration. The goal of using subjective ratings is to capture properties of plural predictability that are not easily captured through n-gram plural predictability from a corpus. As discussed in Chapter 4, because the overt task did not succeed in capturing this nuanced information, future studies could solicit human-based ratings using a methodology more suited to capturing sub-conscious intuitions, such as eye tracking or measuring reaction time. If these methodologies are able to more accurately capture intuitions about morphological predictability, they could be used to test the hypothesis that the context relevant to calculating plural predictability is more than one word. Alternatively, plural probabilities based on larger contexts could be calculated from a larger corpus. However, as discussed in Chapter 4, plural predictability based solely on co-occurrence statistics of word n-grams with plurals may not be the most effective way of capturing plural predictability. It is likely that certain words, such as verbs or adjectives, are more important for predicting plurality than other words. For this reason, measures collected directly from language users may provide more accurate measures of plural predictability.

5.3 Implications and predictions

The findings of this thesis suggest that the biases related to effective communication, namely maintaining a high probability of successful message transmission and maintaining relatively low resource cost, are active at the level of the morpheme. While there are potential alternative motivations for probabilistic reduction (see Jaeger & Buz, 2017), the combined results presented here with regard to both learning and production suggest that the mechanism behind this reduction is not solely production-based. Indeed, these results, in combination with previous results at other levels of linguistic representation, imply that there are overarching biases influencing language at all levels of structure and across learning, production, and perception.

The findings presented in this thesis raise questions about how the influences of these biases interact at different levels of structure to shape language both synchronically and diachronically. Synchronically, how do the influences of word-, segment-, morpheme-, and syntactic structure-level predictability interact? MOP places an emphasis on the message, but how does this extend when there are multiple messages being communicated at once (as is usually the case in language)? And how does this vary when the relative importance of the messages varies? Future studies might examine how production, perception, and learning behavior vary, while modulating predictability at two or more levels of linguistic structure simultaneously. For example, when the predictability of the upcoming syntactic structure is high, is morphological predictability more or less important than when syntactic predictability is low? And can this be manipulated by emphasizing the importance of either the syntactic or morphological message?

Diachronically, MOP makes predictions about how phonological systems are shaped over time. The research presented here, suggesting that both the learning and production of bound morphemes are influenced by predictability, implies that morphological systems may be subject to those same pressures. What implications does this have for language evolution? If learners are biased towards poor learning of predictable morphological material, and producers are biased towards reducing predictable morphological material, how will this be manifested cross-linguistically? This cycle seems to suggest that all redundant morphological material should eventually disappear, yet this is not the case, as language evolution is subject to many other factors. However, the expectation is that, based on these patterns, there would be a tendency for morphological systems to evolve in response to the biases of effective communication. Future studies could explore, cross-linguistically, how plural marking systems balance the demands of transmitting the message of plurality, while conserving resource cost.

6 REFERENCES

Albright, A. (2006). *Similarity calculator*. Retrieved from http://web.mit.edu/albright/www/software/SimilarityCalculator.zip

Aylett, M., & Turk, A. (2006). Language redundancy predicts syllabic duration and the spectral characteristics of vocalic syllable nuclei. *The Journal of the Acoustical Society of America*, *119*(5), 3048–3058.

Baayen, R. H. (2008). *Analyzing linguistic data: A practical introduction to statistics using R*. Cambridge University Press.

Baayen, R. H., Dijkstra, T., & Schreuder, R. (1997). Singulars and plurals in Dutch: Evidence for a parallel dual-route model. *Journal of Memory and Language*, *37*(1), 94–117.

Baayen, R. H., Milin, P., Đurđević, D. F., Hendrix, P., & Marelli, M. (2011). An amorphous model for morphological processing in visual comprehension based on naive discriminative learning. *Psychological Review*, *118*(3), 438.

Baayen, R. H., Piepenbrock, R., & Gulikers, L. (1995). *The CELEX lexical database*. Philadelphia: Linguistic Data Consortium.

Baker, R., Smith, R., & Hawkins, S. (2007). Phonetic differences between mis-and dis-in English prefixed and pseudo-prefixed words. In *Proceedings of the 16th International Congress of Phonetic Sciences*. Citeseer.

Bardovi-Harlig, K. (1992). The use of adverbials and natural order in the development of temporal expression. *IRAL-International Review of Applied Linguistics in Language Teaching*, *30*(4), 299–320.

Bardovi-Harlig, K. (1999). From morpheme studies to temporal semantics. *Studies in Second Language Acquisition*, *21*(03), 341–382.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Bayes, T. (1763). An essay towards solving a problem in the doctrine of chances. *Philisophical Transactions*, *53*, 370–418.

Beckman, M. E., & Edwards, J. (1990). Lengthenings and shortenings and the nature of prosodic constituency. In J. Kingston & M. E. Beckman (Eds.), *Papers in Laboratory Phonology: Volume 1, Between the Grammar and Physics of Speech* (Vol. 1). Cambridge University Press.

Beckner, C., Pierrehumbert, J. B., & Hay, J. (2017). The emergence of linguistic structure in an online iterated learning task. *Journal of Language Evolution*, 1–16. https://doi.org/10.1093/jole/lzx001

Bell, A., Brenier, J. M., Gregory, M., Girand, C., & Jurafsky, D. (2009). Predictability effects on durations of content and function words in conversational English. *Journal of Memory and Language*, *60*(1), 92–111.

Bell, A., Jurafsky, D., Fosler-Lussier, E., Girand, C., & Gildea, D. (1999). Forms of English function words: Effects of disfluencies, turn position, age and sex, and predictability. In *Proceedings of ICPHS-99*. Citeseer.

Bell, A., Jurafsky, D., Fosler-Lussier, E., Girand, C., Gregory, M., & Gildea, D. (2003). Effects of disfluencies, predictability, and utterance position on word form variation in English conversation. *The Journal of the Acoustical Society of America*, *113*(2), 1001–1024.

Berkley, D. M. (1994). The OCP and gradient data. *Studies in the Linguistic Sciences*, *24*(1/2), 59–72.

Bock, K., & Cutting, J. C. (1992). Regulating mental energy: Performance units in language production. *Journal of Memory and Language*, *31*(1), 99–127.

Bock, K., & Miller, C. A. (1991). Broken agreement. *Cognitive Psychology*, *23*(1), 45–93.

Booij, G. (2006). Lexical phonology and morphology.

Borowsky, T. (1987). Antigemination in English phonology. *Linguistic Inquiry*, *18*(4), 671–678.

Brants, T., & Franz, A. (2006). Web 1T 5-gram version 1. Retrieved from http://www.citeulike.org/group/9765/article/4238120

Bresnan, J. (2007). Is syntactic knowledge probabilistic? Experiments with the English dative alternation. *Roots: Linguistics in Search of Its Evidential Base*, 75–96.

Bürki, A., Ernestus, M., Gendrot, C., Fougeron, C., & Frauenfelder, U. H. (2011). What affects the presence versus absence of schwa and its duration: A corpus analysis of French connected speech. *The Journal of the Acoustical Society of America*, *130*(6), 3980–3991.

Bybee, J. L. (2001). *Phonology and language use*. Cambridge: Cambridge University Press.

Bybee, J., & Scheibman, J. (1999). The effect of usage on degrees of constituency: The reduction of don't in English. *Linguistics*, *37*(4), 575–596.

Byrd, D., & Tan, C. C. (1996). Saying consonant clusters quickly. *Journal of Phonetics*, *24*(2), 263–282.

Cambier-Langeveld, G. M. (2000). *Temporal marking of accents and boundaries*. Holland Academic Graphics.

Campbell, W. N., & Isard, S. D. (1991). Segment durations in a syllable frame. *Journal of Phonetics*, *19*(1), 37–47.

Caramazza, A., Laudanna, A., & Romani, C. (1988). Lexical access and inflectional morphology. *Cognition*, *28*(3), 297–332.

Carr, J. W., Smith, K., Cornish, H., & Kirby, S. (2017). The cultural evolution of structured languages in an open-ended, continuous world. *Cognitive Science*, *41*(4), 892–923.

Chateau, D., Knudsen, E. V., & Jared, D. (2002). Masked priming of prefixes and the influence of spelling–meaning consistency. *Brain and Language*, *81*(1), 587–600.

Cheng, P. W., & Holyoak, K. J. (1995). Complex adaptive systems as intuitive statisticians: Causality, contingency, and prediction. In H. L. Roitblat & J.-A. Meyer (Eds.), *Comparative approaches to cognitive science* (pp. 271–302).

Cho, T. (2001). Effects of morpheme boundaries on intergestural timing: Evidence from Korean. *Phonetica*, *58*(3), 129–162.

Cohen, C. (2014). Probabilistic reduction and probabilistic enhancement. *Morphology*, *24*(4), 291–323.

Cohen, C. (2015). Context and paradigms: Two patterns of probabilistic pronunciation variation in Russian agreement suffixes. *The Mental Lexicon*, *10*(3), 313–338.

Cohen Priva, U. (in prep). The interdependence of frequency, predictability, and informativity.

Cohen Priva, U. (2008). Using information content to predict phone deletion. In *Proceedings of the 27th west coast conference on formal linguistics* (pp. 90-98).

Cohen Priva, U. (2012). Sign and signal: Deriving linguistic generalizations from information utility. *Unpublished Doctoral Dissertation, Stanford University*.

Cohen Priva, U. (2015). Informativity affects consonant duration and deletion rates. *Laboratory Phonology*, *6*(2), 243–278.

Crump, M. J., McDonnell, J. V., & Gureckis, T. M. (2013). Evaluating Amazon's Mechanical Turk as a tool for experimental behavioral research. *PloS One*, *8*(3), e57410.

Davis, S. (1991). Coronals and the phonotactics of non-adjacent consonants in English. *The Special Status of Coronals: Internal and External Evidence (Phonetics and Phonology)*, *2*, 49–60.

Dietrich, R., Klein, W., & Noyau, C. (1995). *The acquisition of temporality in a second language* (Vol. 7). John Benjamins Publishing.

Dominguez, A., Alija, M., Rodríguez-Ferreiro, J., & Cuetos, F. (2010). The contribution of prefixes to morphological processing of Spanish words. *European Journal of Cognitive Psychology*, *22*(4), 569–595.

Ellis, N. C. (2006a). Language acquisition as rational contingency learning. *Applied Linguistics*, *27*(1), 1–24.

Ellis, N. C. (2006b). Selective attention and transfer phenomena in L2 acquisition: Contingency, cue competition, salience, interference, overshadowing, blocking, and perceptual learning. *Applied Linguistics*, *27*(2), 164–194.

Erker, D. G. (2010). A subsegmental approach to coda /s/ weakening in Dominican Spanish. *International Journal of the Sociology of Language*, *2010*(203). https://doi.org/10.1515/ijsl.2010.019

Fasold, R. W. (1972). Tense marking in Black English: A linguistic and social analysis. *Urban Language Series*, No. 8. Harcourt College Pub.

Fedzechkina, M., Jaeger, T. F., & Newport, E. L. (2012). Language learners restructure their input to facilitate efficient communication. *Proceedings of the National Academy of Sciences*, *109*(44), 17897–17902.

Fedzechkina, M., Newport, E. L., & Jaeger, T. F. (2017). Balancing effort and information transmission during language acquisition: Evidence from word order and case marking. *Cognitive Science*, *41*(2), 416–446.

Ferrer i Cancho, R. (2005). Zipf's law from a communicative phase transition. *The European Physical Journal B-Condensed Matter and Complex Systems*, *47*(3), 449–457.

Finley, S. (2015). Frequency effects in morpheme segmentation. In *Proceedings of the 37th Annual Conference of the Cognitive Science Society*. Austin, TX: Cognitive Science Society.

Flemming, E. (2004). Contrast and perceptual distinctiveness. In B. Hayes, R. Kirchner, & D. Steriade (Eds.), *Phonetically-based phonology* (pp. 232–276).

Flemming, E. (2010). Modeling listeners: Comments on Pluymaekers et al. and Scarborough. *Laboratory Phonology*, *10*, 587–606.

Ford, M., & Bresnan, J. (2015). Generating data as a proxy for unavailable corpus data: The contextualized sentence completion task. *Corpus Linguistics and Linguistic Theory*, *11*(1), 187–224.

Fosler-Lussier, E., & Morgan, N. (1999). Effects of speaking rate and word frequency on pronunciations in convertional speech. *Speech Communication*, *29*(2), 137–158.

Fougeron, C., & Keating, P. A. (1997). Articulatory strengthening at edges of prosodic domains. *The Journal of the Acoustical Society of America*, *101*(6), 3728–3740.

Franck, J., Vigliocco, G., & Nicol, J. (2002). Subject-verb agreement errors in French and English: The role of syntactic hierarchy. *Language and Cognitive Processes*, *17*(4), 371–404.

Frank, A. F., & Jaeger, T. F. (2008). Speaking rationally: Uniform information density as an optimal strategy for language production. In *Proceedings of the 30th Annual Meeting of the Cognitive Science Society* (pp. 939–944). Cognitive Science Society.

Franks, A. (2011, February 24). Diagnosing collinearity in mixed models from lme4. Retrieved August 24, 2016, from https://hlplab.wordpress.com/2011/02/24/diagnosing-collinearity-in-lme4/

Frisch, S. (1997). *Similarity and frequency in phonology (Unpublished doctoral dissertation)*. Northwestern University.

Frisch, S. A., Pierrehumbert, J. B., & Broe, M. B. (2004). Similarity avoidance and the OCP. *Natural Language & Linguistic Theory*, *22*(1), 179–228.

Gahl, S., Jurafsky, D., & Roland, D. (2004). Verb subcategorization frequencies: American English corpus data, methodological studies, and cross-corpus comparisons. *Behavior Research Methods*, *36*(3), 432–443.

Gahl, S., Yao, Y., & Johnson, K. (2012). Why reduce? Phonological neighborhood density and phonetic reduction in spontaneous speech. *Journal of Memory and Language*, *66*(4), 789–806.

Giacalone Ramat, A. (1995). Function and form of modality in learner Italian. *From Pragmatics to Syntax: Modality in Second Language Acquisition*, *405*, 269.

Gonnerman, L. M., Seidenberg, M. S., & Andersen, E. S. (2007). Graded semantic and phonological similarity effects in priming: Evidence for a distributed connectionist approach to morphology. *Journal of Experimental Psychology: General*, *136*(2), 323.

Gordon, E., Maclagan, M., & Hay, J. (2007). The ONZE corpus. In *Creating and digitizing language corpora* (pp. 82–104). Springer.

Graff, P., & Jaeger, T. (2009). Locality and feature specificity in OCP effects: Evidence from Aymara, Dutch, and Javanese. In *Proceedings from the Annual Meeting of the Chicago Linguistic Society* (Vol. 45, pp. 127–141). Chicago Linguistic Society.

Grainger, J., Colé, P., & Segui, J. (1991). Masked morphological priming in visual word recognition. *Journal of Memory and Language*, *30*(3), 370–384.

Gregory, M. L., Raymond, W. D., Bell, A., Fosler-Lussier, E., & Jurafsky, D. (1999). The effects of collocational strength and contextual predictability in lexical production. In *Chicago Linguistic Society* (Vol. 35, pp. 151–166).

Guy, G. R. (1980). Variation in the group and the individual. In W. Labov (Ed.), *Locating language in time and space* (pp. 1–36). New York: Academic Press.

Guy, G. R. (1991). Explanation in variable phonology: An exponential model of morphological constraints. *Language Variation and Change*, *3*(01), 1–22.

Guy, G. R. (1996). Form and function in linguistic variation. *Towards a Social Science of Language: Papers in Honor of William Labov*, *1*, 221–252.

Guy, G. R., & Boberg, C. (1997). Inherent variability and the obligatory contour principle. *Language Variation and Change*, *9*(02), 149–164.

Hall, K. C., Hume, E., Jaeger, T. F., & Wedel, A. (submitted). The message shapes phonology.

Hanique, I., & Ernestus, M. (2012). The role of morphology in acoustic reduction. *Lingue E Linguaggio*, *11*(2), 147–164.

Hanique, I., Ernestus, M., & Schuppler, B. (2013). Informal speech processes can be categorical in nature, even if they affect many different words. *The Journal of the Acoustical Society of America*, *133*(3), 1644–1655.

Hanique, I., Schuppler, B., & Ernestus, M. (2010). Morphological and predictability effects on schwa reduction: The case of Dutch word-initial syllables. In *11th Annual Conference of the International Speech Communication Association (Interspeech 2010),* (pp. 933–936).

Harrell, F. E. (2001). *Regression modeling strategies: With applications to linear models, logistic regression, and survival analysis*. New York, NY [etc.: Springer.

Harrell, F. E. (2016). Hmisc (Version 4.0-2). Retrieved from https://CRAN.R-project.org/package=Hmisc

Hawkins, S., & Smith, R. (2001). Polysp: A polysystemic, phonetically-rich approach to speech understanding. *Italian Journal of Linguistics*, *13*, 99–188.

Hay, J. (2001). Lexical frequency in morphology: Is everything relative? *Linguistics*, *39*(6; ISSU 376), 1041–1070.

Hay, J. (2004). *Causes and consequences of word structure*. Routledge.

Hay, J., & Baayen, H. (2002). Parsing and productivity. In *Yearbook of morphology 2001* (pp. 203–235). Springer.

Hay, J. B., & Baayen, R. H. (2005). Shifting paradigms: Gradient structure in morphology. *Trends in Cognitive Sciences*, *9*, 342–348.

Hay, J., & Foulkes, P. (2016). The evolution of medial /t/ over real and remembered time. *Language*, *92*(2), 298–330.

Hay, J., Maclagan, M., & Gordon, E. (2008). *New Zealand English*. Edinburgh University Press.

Hay, J., Pierrehumbert, J., & Beckman, M. (2004). Speech perception, well-formedness, and the statistics of the lexicon. *Phonetic Interpretation: Papers in Laboratory Phonology VI, Cambridge University Press, Cambridge*, 58–74.

Hume, E., Rose, D. E., & Spagnol, M. (2014). Maltese Word-initial Singleton-Geminate Contrasts: An Information-theoretic approach. In R. Kager, J.

Grijzenhout, & K. Sebregts (Eds.), *Where the Principles Fail* (pp. 89–101). Utrecht: OTS.

Humphreys, K. R., & Bock, K. (2005). Notional number agreement in English. *Psychonomic Bulletin & Review*, *12*(4), 689–695.

Hundley, J. E. (1987). Functional constraints on plural marker deletion in Peruvian Spanish. *Hispania*, *70*(4), 891. https://doi.org/10.2307/342562

Jaeger, T. F. (2010). Redundancy and reduction: Speakers manage syntactic information density. *Cognitive Psychology*, *61*(1), 23–62.

Jaeger, T. F. (2011). Corpus-based research on language production: Information density and reducible subject relatives. In E. M. Bender & J. E. Arnold (Eds.), *Language from a cognitive perspective: Grammar, usage, and processing. Studies in honor of Tom Wasow* (pp. 161–197). Stanford, CA: CSLI Publications.

Jaeger, T. F. (2013). Production preferences cannot be understood without reference to communication. *Frontiers in Psychology*, *4*.

Jaeger, T. F., & Buz, E. (2017). Signal reduction and linguistic encoding. In *Handbook of Psycholinguistics*. Wiley-Blackwell.

Jaeger, T. F., Furth, K., & Hilliard, C. (2012). Phonological overlap affects lexical selection during sentence production. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*(5), 1439.

Jaeger, T. F., & Norcliffe, E. J. (2009). The cross-linguistic study of sentence production. *Language and Linguistics Compass*, *3*(4), 866–887.

Jurafsky, D., Bell, A., Fosler-Lussier, E., Girand, C., & Raymond, W. D. (1998). Reduction of English function words in switchboard. In *ICSLP-98* (Vol. 7, pp. 3111–3114). Sydney.

Jurafsky, D., Bell, A., & Girand, C. (2002). The role of the lemma in form variation. *Papers in Laboratory Phonology*, *7*, 3–34.

Jurafsky, D., Bell, A., Gregory, M., & Raymond, W. D. (2001). Probabilistic relations between words: Evidence from reduction in lexical production. *Typological Studies in Language*, *45*, 229–254.

Kemps, R. J., Ernestus, M., Schreuder, R., & Baayen, R. H. (2005). Prosodic cues for morphological complexity: The case of Dutch plural nouns. *Memory & Cognition*, *33*(3), 430–446.

Kemps, R. J., Wurm, L. H., Ernestus, M., Schreuder, R., & Baayen, R. H. (2005). Prosodic cues for morphological complexity in Dutch and English. *Language and Cognitive Processes*, *20*(1-2), 43–73.

Kidd, C., Piantadosi, S. T., & Aslin, R. N. (2014). The Goldilocks effect in infant auditory attention. *Child Development*, *85*(5), 1795–1804.

Kim, J. (2016). Perceptual associations between words and speaker age. *Laboratory Phonology: Journal of the Association for Laboratory Phonology*, *7*(1), 18.

Kirby, S., Cornish, H., & Smith, K. (2008). Cumulative cultural evolution in the laboratory: An experimental approach to the origins of structure in human language. *Proceedings of the National Academy of Sciences*, *105*(31), 10681–10686.

Kirov, C., & Wilson, C. (2013). Bayesian speech production: Evidence from latency and hyperarticulation. In *Proceedings of the 35th Annual Meeting of the Cognitive Science Society of America* (pp. 788–793).

Klatt, D. H. (1976). Linguistic uses of segmental duration in English: Acoustic and perceptual evidence. *The Journal of the Acoustical Society of America*, *59*(5), 1208–1221.

Kravtchenko, E. (2014). Predictability and syntactic production: Evidence from subject omission in Russian. In *Proceedings of the 36th Annual Meeting of the Cognitive Science Society of America*. Austin, TX: Cognitive Science Society.

Kristiansen, T. (2009). The macro-level social meanings of late-modern Danish accents. *Acta Linguistica Hafniensia*, *41*(1), 167–192.

Kristiansen, T. (2010). Conscious and subconscious attitudes towards English influence in the Nordic countries: evidence for two levels of language ideology. *International Journal of the Sociology of Language*, *2010*(204), 59–95.

Kruschke, J. K., & Blair, N. J. (2000). Blocking and backward blocking involve learned inattention. *Psychonomic Bulletin & Review*, *7*(4), 636–645.

Kuperman, V., Pluymaekers, M., Ernestus, M., & Baayen, R. H. (2007). Morphological predictability and acoustic duration of interfixes in Dutch compounds. *Journal of the Acoustical Society of America*, *121*(4), 2261–2271.

Kuperman, V., & Van Dyke, J. A. (2013). Reassessing word frequency as a determinant of word recognition for skilled and unskilled readers. *Journal of Experimental Psychology: Human Perception and Performance*, *39*(3), 802.

Kurumada, C., & Grimm, S. (2017). Communicative efficiency in language production and learning: Optional plural marking. *Proceeedings of the 39th Annual Meeting of the Cognitive Science Society*.

Kurumada, C., & Jaeger, T. F. (2015). Communicative efficiency in language production: Optional case-marking in Japanese. *Journal of Memory and Language*, *83*, 152–178.

Labov, W. (1968). *A study of the non-standard English of Negro and Puerto Rican speakers in New York City* (Vol. 1). Columbia University.

Labov, W. (1972). *Sociolinguistic patterns*. University of Pennsylvania Press.

Laplace, P.-S. (1812). *Théorie analytique des probabilités*.

Laudanna, A., Burani, C., & Cermele, A. (1994). Prefixes as processing units. *Language and Cognitive Processes*, *9*(3), 295–316.

Leben, W. R. (1973). *Suprasegmental phonology.* Massachusetts Institute of Technology.

Leek, M. R. (2001). Adaptive procedures in psychophysical research. *Attention, Perception, & Psychophysics*, *63*(8), 1279–1292.

Lehiste, I. (1970). Temporal organization of spoken language. *Ohio State University Working Papers*.

Lombard, E. (1910a). A propos de la note rectificative et de la rectification. *Annales Des Maladies de l'Oreille et Du Larynx*, *36*, 111–112.

Lombard, E. (1910b). Note rectificative. *Annales Des Maladies de l'Oreille et Du Larynx*, *36*, 34–35.

Lombard, E. (2011). Le signe de l'élévation de la voix. *Annales Des Maladies de l'Oreille et Du Larynx*, *37*, 101–119.

Losiewicz, B. L. (1992). *The effect of frequency on linguistic morphology*. University of Texas at Austin.

MacWhinney, B. (1997). Second language acquisition and the competition model. In A. M. B. de Groot & J. F. Kroll (Eds.), *Tutorials in bilingualism: Psycholinguistic perspectives* (pp. 113–142).

Marslen-Wilson, W., Tyler, L. K., Waksler, R., & Older, L. (1994). Morphology and meaning in the English mental lexicon. *Psychological Review*, *101*(1), 3.

McCarthy, J. J. (1986). OCP effects: Gemination and antigemination. *Linguistic Inquiry*, 207–263.

Melnick, R., Jaeger, T. F., & Wasow, T. (2010). *Speakers employ fine-grained probabilistic knowledge*. Poster presentation presented at the Conference on Human Sentence Processsing, NYU.

Miesel, J. M. (1987). Reference to past events and actions in the development of natural language acquisition. *First and Second Language Acquisition Processes*, 206–224.

Mousikou, P., Strycharczuk, P., Turk, A., Rastle, K., & Scobbie, J. M. (2015). Morphological effects on pronunciation. *Proceedings of the 18th ICPhS*, (0816).

Norcliffe, E., & Jaeger, T. F. (2016). Predicting head-marking variability in Yucatec Maya relative clause production. *Language and Cognition*, *8*(02), 167–205.

Odden, D. (1986). On the role of the Obligatory Contour Principle in phonological theory. *Language*, *62*(2), 353. https://doi.org/10.2307/414677

O'Donnell, T. J. (2015). *Productivity and reuse in language: A theory of linguistic computation and storage*. MIT Press.

Pate, J. K., & Goldwater, S. (2015). Talkers account for listener and channel characteristics to communicate efficiently. *Journal of Memory and Language*, *78*, 1–17.

Pavlov, I. P. (1927). *Conditioned Reflexes. An Investigation of the Physiological Activity of the Cerebral Cortex... Translated and Edited by GV Anrep*. (G. V. ANREP, Trans.). London.

Piantadosi, S. T., Tily, H., & Gibson, E. (2011). Word lengths are optimized for efficient communication. *PNAS*, *108*(9), 3526–3529.

Pierce, J. R. (1980). *An introduction to Information Theory: Symbols, signals, and noise.* (2nd ed.). New York, NY: Dover Publications.

Pierrehumbert, J. (1993). Dissimilarity in the Arabic verbal roots. *Proceedings of the 23rd Meeting of the Northeastern Linguistic Society, Graduate Student Association, U. Mass. Amherst*, 367–381.

Pierrehumbert, J. (1994). Syllable structure and word structure: A study of triconsonantal clusters in English. *Papers in Laboratory Phonology*, *3*.

Plag, I., Homann, J., & Kunter, G. (2017). Homophony and morphology: The acoustics of word-final S in English. *Journal of Linguistics*, *53*(1), 181–216.

Pluymaekers, M., Ernestus, M., & Baayen, R. (2005a). Articulatory planning is continuous and sensitive to informational redundancy. *Phonetica*, *62*(2-4), 146–159.

Pluymaekers, M., Ernestus, M., & Baayen, R. H. (2005b). Lexical frequency and acoustic reduction in spoken Dutch. *The Journal of the Acoustical Society of America*, *118*(4), 2561. https://doi.org/10.1121/1.2011150

Pluymaekers, M., Ernestus, M., Baayen, R. H., & Booij, G. (2010). Morphological effects on fine phonetic detail: The case of Dutch-igheid. *Laboratory Phonology*, *10*, 511–531.

Poplack, S. (1980). Deletion and disambiguation in Puerto Rican Spanish. *Language*, *56*(2), 371. https://doi.org/10.2307/413761

Rácz, P., Hay, J. B., & Pierrehumbert, J. B. (2017). Social salience discriminates learnability of contextual cues in an artificial language. *Frontiers in Psychology*, *8*, 51.

Ramscar, M., Dye, M., & Klein, J. (2013). Children value informativity over logic in word learning. *Psychological Science*, *24*(6), 1017–1023.

Ramscar, M., Dye, M., & McCauley, S. M. (2013). Error and expectation in language learning: The curious absence of mouses in adult speech. *Language*, *89*(4), 760–793.

Raymond, W., Dautricourt, R., & Hume, E. (2006). Word-medial /t,d/ deletion in spontaneous speech: Modeling the effects of extra-linguistic, lexical, and phonological factors. *Language Variation and Change*, *18*, 55–97.

R Core Team. (2015). *R: A language and environment for statistical computing.* Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org/

Rescorla, R. A. (1969). Conditioned inhibition of fear. In W. K. Honig & N. J. Mackintosh (Eds.), *Fundamental issues in associative learning* (pp. 65–89). Halifax: Dalhousie University Press.

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. Black & W. F. Prokasy (Eds.), *Classical Conditioning II*. New York: Appleton-Century-Crofts.

Schumacher, R. A., Pierrehumbert, J. B., & LaShell, P. (2014). Reconciling inconsistency in encoded morphological distinctions in an artificial language. In *Proceedings of the 36th Annual Meeting of the Cognitive Science Society of America*. Austin, TX: Cognitive Science Society.

Schuppler, B., van Dommelen, W. A., Koreman, J., & Ernestus, M. (2012). How linguistic and probabilistic properties of a word affect the realization of its final /t/: Studies at the phonemic and sub-phonemic level. *Journal of Phonetics*, *40*(4), 595–607.

Seyfarth, S. (2014). Word informativity influences acoustic duration: Effects of contextual predictability on lexical representation. *Cognition*, *133*(1), 140–155.

Seyfarth, S. J. (2016). *Contextual and morphological effects in speech production*. University of California, San Diego.

Seyfarth, S., Buz, E., & Jaeger, T. F. (2016). Dynamic hyperarticulation of coda voicing contrasts. *The Journal of the Acoustical Society of America*, *139*(2), EL31–EL37.

Shannon, C. E. (1948). A Mathematical Theory of Communication. *Mobile Computating and Communications Review*, *5*(1), 3–55.

Solomon, E. S., & Pearlmutter, N. J. (2004). Semantic integration and syntactic planning in language production. *Cognitive Psychology*, *49*(1), 1–46.

Song, J. Y., Demuth, K., Shattuck-Hufnagel, S., & Ménard, L. (2013). The effects of coarticulation and morphological complexity on the production of English coda clusters: Acoustic and articulatory evidence from 2-year-olds and adults using ultrasound. *Journal of Phonetics*, *41*(3), 281–295.

Sugahara, M., & Turk, A. (2009). Durational correlates of English sublexical constituent structure. *Phonology*, *26*(3), 477–524. https://doi.org/10.1017/S0952675709990248

Tassoni, C. J. (1995). The least mean squares network with information coding: A model of cue learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*(1), 193.

Thornton, R., & MacDonald, M. C. (2003). Plausibility and grammatical agreement. *Journal of Memory and Language*, *48*(4), 740–759.

Tily, H., & Piantadosi, S. (2009). Refer efficiently: Use less informative expressions for more predictable meanings. In *Proceedings of the workshop on the production of referring expressions: Bridging the gap between computational and empirical approaches to reference* (pp. 1–8).

Torreira, F., & Ernestus, M. (2012). Weakening of intervocalic /s/ in the Nijmegen Corpus of Casual Spanish. *Phonetica*, *69*(3), 124–148. https://doi.org/10.1159/000343635

Toscano, J. C., Buxó-Lugo, A., & Watson, D. G. (2015). Using game-based approaches to increase level of engagement in research and education. In *Teachercraft* (pp. 139–151). ETC Press.

van Son, R. J. J. H., & Pols, L. C. W. (2003). How efficient is speech? In E. H. Berkman (Ed.), *Proceedings of the Institute of Phonetic Sciences*. Amsterdam.

Viebahn, M. C., Ernestus, M., & McQueen, J. M. (2012). Co-occurrence of reduced word forms in natural speech. In *INTERSPEECH 2012: 13th Annual Conference of the International Speech Communication Association* (pp. 2019–2022).

von Ahn, L. (2006). Games with a purpose. *IEEE Computer*, *39*(6), 92–94.

Wagner, A. R. (1969). Stimulus validity and stimulus selection. In W. K. Honig & N. J. Mackintosh (Eds.), *Fundamental issues in associative learning* (pp. 90–122). Halifax: Dalhousie University Press.

Wagner, A. R. (1970). Stimulus selection and a "Modified Continuity Theory." *Psychology of Learning and Motivation*, *3*, 1–41.

Walker, A., & Hay, J. (in prep). Congruence between "word gender" and "voice gender" also facilitates lexical access.

Walker, A., & Hay, J. (2011). Congruence between "word age" and "voice age" facilitates lexical access. *Laboratory Phonology*, *2*(1). https://doi.org/10.1515/labphon.2011.007

Walsh, T., & Parker, F. (1983). The duration of morphemic and non-morphemic /s/ in English. *Journal of Phonetics*, *11*, 201–206.

Wang, S., Huang, C. R., Yao, Y., & Chan, A. (2014). Exploring mental lexicon in an efficient and economic way: Crowdsourcing method for linguistic experiments. *COLING 2014*, 105-113.

Wasow, T., Jaeger, T. F., & Orr, D. (2011). Lexical variation in relativizer frequency. In H. Simon & H. Wiese (Eds.), *Expecting the unexpected: Exceptions in grammar* (pp. 175–195).

Yip, M. (1988). The Obligatory Contour Principle and phonological rules: A loss of identity. *Linguistic Inquiry*, *19*(1), 65–100.

Zimmermann, J. (2016). Morphological status and acoustic realization: Findings from NZE. In C. Carignan & M. D. Tyler (Eds.), *Proceedings of the Sixteenth Australasian International Conference on Speech Science and Technology (SST-2016)* (pp. 201–204). Canberra: ASSTA.

Zipf, G. K. (1949). *Human behavior and the principle of least effort*.

Zue, V. W., & Laferrière, M. (1979). Acoustic study of medial/t, d/in American English. *The Journal of the Acoustical Society of America*, *66*(4), 1039–1050.

Zuur, A. F., Ieno, E. N., & Elphick, C. S. (2010). A protocol for data exploration to avoid common statistical problems. *Methods in Ecology and Evolution*, *1*(1), 3–14.

7 APPENDICES

APPENDIX 1: STIMULI FOR STUDY 1

	n-medial				l-medial			
	Singular	**Plural**	**Option 3**	**Option 4**	**Singular**	**Plural**	**Option 3**	**Option 4**
Cue A	banop	bannop	banopp	bannopp	balop	ballop	balopp	ballopp
	panop	pannop	panopp	pannopp	palop	pallop	palopp	pallopp
	danop	dannop	danopp	dannopp	dalop	dallop	dalopp	dallopp
	kanop	kannop	kanopp	kannopp	kalop	kallop	kalopp	kallopp
	branop	brannop	branopp	brannopp	bralop	brallop	bralopp	brallopp
	pranop	prannop	pranopp	prannopp	pralop	prallop	pralopp	prallopp
	dranop	drannop	dranopp	drannopp	dralop	drallop	dralopp	drallopp
	tranop	trannop	tranopp	trannopp	tralop	trallop	tralopp	trallopp
	granop	grannop	granopp	grannopp	gralop	grallop	gralopp	grallopp
	blanop	blannop	blanopp	blannopp	blalop	blallop	blalopp	blallopp
	planop	plannop	planopp	plannopp	plalop	plallop	plalopp	plallopp
	dwanop	dwannop	dwanopp	dwannopp	dwalop	dwallop	dwalopp	dwallopp
	twanop	twannop	twanopp	twannopp	twalop	twallop	twalopp	twallopp
	glanop	glannop	glanopp	glannopp	glalop	glallop	glalopp	glallopp
	klanop	klannop	klanopp	klannopp	klalop	klallop	klalopp	klallopp
	fanop	fannop	fanopp	fannopp	falop	fallop	falopp	fallopp
	vanop	vannop	vanopp	vannopp	valop	vallop	valopp	vallopp
	thanop	thannop	thanopp	thannopp	thalop	thallop	thalopp	thallopp
	zanop	zannop	zanopp	zannopp	zalop	zallop	zalopp	zallopp
	thranop	thrannop	thranopp	thrannopp	thralop	thrallop	thralopp	thrallopp
	shranop	shrannop	shranopp	shrannopp	shralop	shrallop	shralopp	shrallopp
	franop	frannop	franopp	frannopp	fralop	frallop	fralopp	frallopp
	kwanop	kwannop	kwanopp	kwannopp	kwalop	kwallop	kwalopp	kwallopp
	swanop	swannop	swanopp	swannopp	swalop	swallop	swalopp	swallopp
Cue B	bapol	bapoll	bappol	bappoll	bapon	baponn	bappon	bapponn
	papol	papoll	pappol	pappoll	papon	paponn	pappon	papponn
	dapol	dapoll	dappol	dappoll	dapon	daponn	dappon	dapponn
	kapol	kapoll	kappol	kappoll	kapon	kaponn	kappon	kapponn
	brapol	brapoll	brappol	brappoll	brapon	braponn	brappon	brapponn
	prapol	prapoll	prappol	prappoll	prapon	praponn	prappon	prapponn
	drapol	drapoll	drappol	drappoll	drapon	draponn	drappon	drapponn
	trapol	trapoll	trappol	trappoll	trapon	traponn	trappon	trapponn
	grapol	grapoll	grappol	grappoll	grapon	graponn	grappon	grapponn
	blapol	blapoll	blappol	blappoll	blapon	blaponn	blappon	blapponn
	plapol	plapoll	plappol	plappoll	plapon	plaponn	plappon	plapponn
	dwapol	dwapoll	dwappol	dwappoll	dwapon	dwaponn	dwappon	dwapponn
	twapol	twapoll	twappol	twappoll	twapon	twaponn	twappon	twapponn
	glapol	glapoll	glappol	glappoll	glapon	glaponn	glappon	glapponn
	klapol	klapoll	klappol	klappoll	klapon	klaponn	klappon	klapponn
	fapol	fapoll	fappol	fappoll	fapon	faponn	fappon	fapponn
	vapol	vapoll	vappol	vappoll	vapon	vaponn	vappon	vapponn
	thapol	thapoll	thappol	thappoll	thapon	thaponn	thappon	thapponn
	zapol	zapoll	zappol	zappoll	zapon	zaponn	zappon	zapponn
	thrapol	thrapoll	thrappol	thrappoll	thrapon	thraponn	thrappon	thrapponn
	shrapol	shrapoll	shrappol	shrappoll	shrapon	shraponn	shrappon	shrapponn
	frapol	frapoll	frappol	frappoll	frapon	fraponn	frappon	frapponn
	kwapol	kwapoll	kwappol	kwappoll	kwapon	kwaponn	kwappon	kwapponn
	swapol	swapoll	swappol	swappoll	swapon	swaponn	swappon	swapponn

	n-medial				l-medial			
	Singular	**Plural**	**Option 3**	**Option 4**	**Singular**	**Plural**	**Option 3**	**Option 4**
Cues A + B	banol	bannoll	bannol	banoll	balon	ballonn	ballon	balonn
	panol	pannoll	pannol	panoll	palon	pallonn	pallon	palonn
	danol	dannoll	dannol	danoll	dalon	dallonn	dallon	dalonn
	kanol	kannoll	kannol	kanoll	kalon	kallonn	kallon	kalonn
	branol	brannoll	brannol	branoll	bralon	brallonn	brallon	bralonn
	pranol	prannoll	prannol	pranoll	pralon	prallonn	prallon	pralonn
	dranol	drannoll	drannol	dranoll	dralon	drallonn	drallon	dralonn
	tranol	trannoll	trannol	tranoll	tralon	trallonn	trallon	tralonn
	granol	grannoll	grannol	granoll	gralon	grallonn	grallon	gralonn
	blanol	blannoll	blannol	blanoll	blalon	blallonn	blallon	blalonn
	planol	plannoll	plannol	planoll	plalon	plallonn	plallon	plalonn
	dwanol	dwannoll	dwannol	dwanoll	dwalon	dwallonn	dwallon	dwalonn
	twanol	twannoll	twannol	twanoll	twalon	twallonn	twallon	twalonn
	glanol	glannoll	glannol	glanoll	glalon	glallonn	glallon	glalonn
	klanol	klannoll	klannol	klanoll	klalon	klallonn	klallon	klalonn
	fanol	fannoll	fannol	fanoll	falon	fallonn	fallon	falonn
	vanol	vannoll	vannol	vanoll	valon	vallonn	vallon	valonn
	thanol	thannoll	thannol	thanoll	thalon	thallonn	thallon	thalonn
	zanol	zannoll	zannol	zanoll	zalon	zallonn	zallon	zalonn
	thranol	thrannoll	thrannol	thranoll	thralon	thrallonn	thrallon	thralonn
	shranol	shrannoll	shrannol	shranoll	shralon	shrallonn	shrallon	shralonn
	franol	frannoll	frannol	franoll	fralon	frallonn	frallon	fralonn
	kwanol	kwannoll	kwannol	kwanoll	kwalon	kwallonn	kwallon	kwalonn
	swanol	swannoll	swannol	swanoll	swalon	swallonn	swallon	swalonn

APPENDIX 2: INSTRUCTIONS FOR STUDY 3

Purpose

This task is part of a research project at the University of Canterbury in Christchurch, New Zealand. For more information, please click the link below:

link here

Overview

In this task we want your help evaluating whether a plural or singular word sounds better after the given context.

The words you can choose are not real words, but they look like real words. For example, 'wug' is meant to represent a singular word and 'wugs' is meant to represent a plural word. A 'wug' could be a concrete noun, as in "I see five wugs outside" or "look at that wug". It could also be more abstract, as in "the very thought of it fills me with wug".

When you see a context, just pick which one you think sounds better. Use the scale to indicate how sure you are of your answer.

These contexts are drawn from real speech, so may or may not be complete sentences. They may include stutters (eg. "t- today") or words like "um" or "ah", which indicate a pause. They may also include names, but all of the names have been changed. Please try to think of these sentences as spoken language, and just choose the word you think fits best.

We Provide

- Content (the context preceding the word)
- A way for you to choose which word you think fits best.

Thank You!

Thank you very much for your work!

APPENDIX 3: INFORMATION SHEET FOR STUDY 3

Linguistics Department,
School of Language, Social and Political Sciences
Email: darcy.rose@pg.canterbury.ac.nz

Choosing singular or plural words based on context

Information Sheet for Participants

I am Darcy Rose, a researcher at the University of Canterbury, and this research project is a part of my PhD thesis. The purpose of this project is to study how likely a plural noun is to occur after a given string of words in English.

Your involvement in this project will involve reading sequences of words and deciding whether the string of words would be best followed by a singular or plural noun. You will be asked to indicate how sure you are of your choice by choosing a point on a ten-point scale between the singular and plural. Your responses will be recorded by CrowdFlower and returned to me, the researcher. CrowdFlower will also provide me with your worker identification number and information on your approximate location, based on your IP address. You may complete as many pages of the task as you wish, and you will be compensated for each page you complete.

There are no known risks to you in the performance of the tasks asked of you in this study.

You may receive a copy of the project results by contacting the researcher. If you would like to receive a copy of the results, please contact the researcher via the above contact information.

Participation in this study is voluntary and you have the right to withdraw at any stage without penalty. If you would like to withdraw, please notify the researcher and any information related to you will be removed.

The results of the project may be published, but you may be assured of the complete confidentiality of data gathered in this investigation: we will not collect any information about you except your approximate location and a worker identification number. In any published reports, you will be assigned a subject number which is different from your worker identification number in order to ensure anonymity and confidentiality. Your worker identification number will be securely stored in a password-protected computer, to which only the researcher and supervisor will have access. Your responses will be stored on the server of the New Zealand Institute of Language, Brain and Behaviour, and may be used in future studies and/or made publically available as part of an online appendix to a journal article.

The project is being carried out as a requirement for the degree of Doctor of Philosophy in Linguistics at the University Canterbury by Darcy Rose, under the supervision of Professor Beth Hume, who can be contacted at beth.hume@canterbury.ac.nz. She will be pleased to discuss any concerns you may have about participation in the project.

This project has been reviewed and approved by the University of Canterbury Human Ethics Committee, and participants should address any complaints to The Chair, Human Ethics Committee, University of Canterbury, Private Bag 4800, Christchurch (human-ethics@canterbury.ac.nz).

Darcy Rose

APPENDIX 4: SAMPLE OF 1-WORD CONTEXTS FOR STUDY 3

few	boy	cultural	probably	heads
eighteen	street	scottish	throw	get
and	cause	may	in	told
the	finding	western	tea	lovely
two	thirty	seven	concrete	father's
of	eleven	engineering	or	his
six	were	us	thousand	through
high	as	seeing	certain	you
with	most	gradually	off	her
by	some	various	making	w~
be	funny	out	bucket	cigarettes
on	same	four	ear	museum
any	these	stream	machine	fast
at	other	put	huge	young
early	our	inch	sitting	to
a	leading	side	more	over
big	different	hundred	first	n
those	their	nick	landing	made
three	short	twice	type	ten
girl	doing	forty	ones	say
just	blimmin	little	burst	ah
thirteen	married	real	tremendous	them
both	full	ahh	mud	five
small	for	playing	based	nine
female	worked	practically	bridges	long
had	planned	one	chemist	great
your	twenty	bad	varsity	hydraulic
rabbit	do	grab	that	poetry
good	close	several	about	fifty
everyday	gathering	mixed	taken	only
from	paper	sheep	where	many
twelve	silly	cattle	an	hot
eight	copy	cold	spare	best
odd	usual	saw	play	s
are	felt	macetown	old	practical
know	straw	er	wet	rock
all	sun	worst	half	perhaps
no	pot	walking	make	buying
heavy	read	clear	saturday	we'd
church	my	similar	floor	beautiful

APPENDIX 5: SAMPLE OF 5-WORD CONTEXTS FOR STUDY 3

lot more about different place
personally he'd never admit that
we were you know really
longer um they sent in
me it's too advanced in
then there was my sister
the whole lot or just
but ah nothing much about
ah a company to give
six year olds learn about
and vegetables and fruit and
ever talk to them or
horses up the orchards and
um freezing cold umbrellas and
and wicks at good and
the sandhills played chasing and
of fruit and vegetables and
as in visiting huts and
she got beaten up and
made with fern trees and
farewells and lots of welcome
that's come on tv for
i think you've got harder
you see total strangers having
but it was just christmas
were no um very bad
to make all our winter
like john smithers who had
the easy pay system for
too small f~ s~ with
picked two milk bucketfuls of
run and those sort of
know make a poster of
to do you could buy
always remember the lot of
ah tui did lots of
um just a couple of

and he was only playing
the first day with his
they're still quite frisky n
he you know all his
he got all the old
four shots which were reasonable
and i was at alice
fines you know the big
would be in the old
ships they were what big
there one of the model
and he built his own
making an appointment with my
she's even got her red
my book or play computer
little points like she's making
and ahh most of my
just it's just the little
can see all the sand
was our one of our
them ah dictation so different
home that's open to their
catch the eels in their
um one of those boxing
er roller skating and other
a crystal and having ear
to get it right other
of things they tried different
why don't you wear your
it and even then your
stay in hospital for five
so if there's too many
done so i got two
had sort of like two
they get holidays for two
i think they had two
from you after so many